

CONVERGENCE AND CONVERGENCE RATE OF STOCHASTIC GRADIENT SEARCH IN THE CASE OF MULTIPLE AND NON-ISOLATED EXTREMA

VLADISLAV B. TADIĆ *

Abstract. The asymptotic behavior of stochastic gradient algorithms is studied. Relying on some results of differential geometry (Lojasiewicz gradient inequality), the almost sure point-convergence is demonstrated and relatively tight almost sure bounds on the convergence rate are derived. In sharp contrast to all existing result of this kind, the asymptotic results obtained here do not require the objective function (associated with the stochastic gradient search) to have an isolated minimum at which the Hessian of the objective function is strictly positive definite. Using the obtained results, the asymptotic behavior of recursive prediction error identification methods is analyzed. The convergence and convergence rate of supervised learning algorithms are also studied relying on these results.

Key words. Stochastic gradient search, point-convergence, convergence rate, Lojasiewicz gradient inequality, system identification, recursive prediction error, ARMA models, machine learning, supervised learning, feedforward neural networks.

AMS subject classifications. Primary 62L20; Secondary 90C15, 93E12, 93E35.

1. Introduction. Stochastic optimization is at the core of many engineering, statistics and finance problems. A stochastic optimization problem can be described as the minimization (or maximization) of an objective function in a situation when only noise-corrupted observations of the function values are available. Such a problem can be solved efficiently by stochastic gradient search, a stochastic approximation version of the deterministic steepest descent method. Due to its excellent performance (generality, robustness, low complexity, easy implementation), stochastic gradient algorithms have gained a wide attention in the literature and have found a broad range of applications in diverse areas such as signal processing, system identification, automatic control, machine learning, operations research, statistical inference, econometrics and finance (see e.g. [2], [7], [9], [10], [11], [16], [17], [22], [24], [25], [26] and reference cited therein).

Various asymptotic properties of stochastic gradient algorithms have been the subject of a number of papers and books (see [1], [14], [16], [24], [26] and references cited therein). Among them, the almost sure convergence and the convergence rate have received the greatest attention, as these properties most precisely characterize the asymptotic behavior and efficiency of stochastic gradient search. Although the existing results provide a good insight into the convergence and convergence rate, they hold only under very restrictive conditions. More specifically, the existing results require the objective function (which the stochastic gradient search minimizes) to have an isolated minimum such that the Hessian of the objective function is strictly positive definite at the minimum and such that the attraction domain of the minimum is infinitely often visited by the algorithm iterates. However, in the case of complex, high-dimensional high-nonlinear algorithms, this is not only hard (if possible at all) to verify, but is likely not to be true.

In this paper, the convergence and convergence rate of stochastic gradient search are analyzed when the objective function has multiple non-isolated minima (notice

*Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, United Kingdom. (v.b.tadic@bristol.ac.uk).

that at a non-isolated minimum, the Hessian can be semi-definite at best). Using some results of differential geometry (Łojasiewicz gradient inequality), the almost sure point-convergence is demonstrated and relatively tight almost sure bounds on the convergence rate are derived. The obtained results cover a wide class of complex stochastic gradient algorithms. We show how they can be used to analyze the asymptotic behavior of recursive prediction error algorithms for identification of linear stochastic systems. We also show how the convergence and convergence rate of supervised learning in feedforward neural networks can be analyzed using the results obtained here.

The paper is organized as follows. In Section 2, stochastic gradient algorithms with additive noise are considered and the main results of the paper are presented. Section 3 is devoted to stochastic gradient algorithms with Markovian dynamics. Sections 4 and 5 contain examples of the results reported in Sections 2 and 3. In Section 4, supervised learning algorithms for feedforward neural networks are studied, while recursive prediction error algorithms for identification of linear stochastic systems are analyzed in Section 5. Sections 6 – 9 contain the proofs of the results presented in Sections 2 – 5.

2. Main Results. In this section, the convergence and convergence rate of the following algorithm is analyzed:

$$\theta_{n+1} = \theta_n - \alpha_n(\nabla f(\theta_n) + \xi_n), \quad n \geq 0. \quad (2.1)$$

Here, $f : \mathbb{R}^{d_\theta} \rightarrow \mathbb{R}$ is a differentiable function, while $\{\alpha_n\}_{n \geq 0}$ is a sequence of positive real numbers. θ_0 is an \mathbb{R}^{d_θ} -valued random variable defined on a probability space (Ω, \mathcal{F}, P) , while $\{\xi_n\}_{n \geq 0}$ is an \mathbb{R}^{d_θ} -valued stochastic process defined on the same probability space. To allow more generality, we assume that for each $n \geq 0$, ξ_n is a random function of $\theta_0, \dots, \theta_n$. In the area of stochastic optimization, recursion (2.1) is known as a stochastic gradient search (or stochastic gradient algorithm), while function $f(\cdot)$ is referred to as an objective function. For further details see [22], [26] and references given therein.

Throughout the paper, unless otherwise stated, the following notation is used. The Euclidean norm is denoted by $\|\cdot\|$, while $d(\cdot, \cdot)$ stands for the distance induced by the Euclidean norm. S is the sets of stationary points of $f(\cdot)$, i.e.,

$$S = \{\theta \in \mathbb{R}^{d_\theta} : \nabla f(\theta) = 0\}.$$

Sequence $\{\gamma_n\}_{n \geq 0}$ is defined by $\gamma_0 = 0$ and

$$\gamma_n = \sum_{i=0}^{n-1} \alpha_i$$

for $n \geq 1$. For $t \in (0, \infty)$ and $n \geq 0$, $a(n, t)$ is the integer defined as

$$a(n, t) = \max \{k \geq n : \gamma_k - \gamma_n \leq t\}.$$

Algorithm (2.1) is analyzed under the following assumptions:

ASSUMPTION 2.1. $\lim_{n \rightarrow \infty} \alpha_n = 0$ and $\sum_{n=0}^{\infty} \alpha_n = \infty$.

ASSUMPTION 2.2. *There exists a real number $r \in (1, \infty)$ such that*

$$\xi = \limsup_{n \rightarrow \infty} \max_{n \leq k < a(n, 1)} \left\| \sum_{i=n}^k \alpha_i \gamma_i^r \xi_i \right\| < \infty$$

w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$.

ASSUMPTION 2.3. For any compact set $Q \subset \mathbb{R}^{d_\theta}$ and any $a \in f(Q)$, there exist real numbers $\delta_{Q,a} \in (0, 1]$, $\mu_{Q,a} \in (1, 2]$, $M_{Q,a} \in [1, \infty)$ such that

$$|f(\theta) - a| \leq M_{Q,a} \|\nabla f(\theta)\|^{\mu_{Q,a}} \quad (2.2)$$

for all $\theta \in Q$ satisfying $|f(\theta) - a| \leq \delta_{Q,a}$.

REMARK 2.1. As an immediate consequence of Assumption 2.3, we have that for each $\theta \in R^{d_\theta}$, there exist real numbers $\delta_\theta \in (0, 1]$, $\mu_\theta \in (1, 2]$, $M_\theta \in [1, \infty)$ such that

$$|f(\theta') - f(\theta)| \leq M_\theta \|\nabla f(\theta')\|^{\mu_\theta} \quad (2.3)$$

for all $\theta' \in \mathbb{R}^{d_\theta}$ satisfying $\|\theta' - \theta\| \leq \delta_\theta$. If $\theta \in S$, μ_θ and M_θ can be selected as

$$\mu_\theta = (1 - \varepsilon) \liminf_{\theta' \rightarrow \theta} \frac{\log |f(\theta') - f(\theta)|}{\log \|\nabla f(\theta')\|}, \quad M_\theta = (1 + \varepsilon) \limsup_{\theta' \rightarrow \theta} \frac{|f(\theta') - f(\theta)|}{\|\nabla f(\theta')\|^{\mu_\theta}}$$

where ε is a small positive constant (since $\{\theta_n\}_{n \geq 0}$ converges to S , the values of μ_θ , M_θ for $\theta \notin S$ are not relevant to the problems studied in the paper). Moreover, if $Q \subseteq \{\theta' \in R^{d_\theta} : \|\theta' - \theta\| \leq \delta_\theta\}$ and $a = f(\theta) \in Q$ for some $\theta \in R^{d_\theta}$, $\mu_{Q,a}$ and $M_{Q,a}$ can be selected as $\mu_{Q,a} = \mu_\theta$, $M_{Q,a} = M_\theta$.

REMARK 2.2. In order for Assumption 2.3 to be true, it is quite sufficient that the assumption holds locally in an open vicinity of S , i.e., that there exists an open set $V \supset S$ with the following property: For any compact set $Q \subset V$ and any $a \in f(Q)$, there exist real numbers $\delta_{Q,a} \in (0, 1]$, $\mu_{Q,a} \in (1, 2]$, $M_{Q,a} \in [1, \infty)$ such that (2.2) holds for all $\theta \in Q$ satisfying $|f(\theta) - a| \leq \delta_{Q,a}$ (see Appendix for details).

Assumption 2.1 correspond to the sequence $\{\alpha_n\}_{n \geq 0}$ and is widely used in the asymptotic analysis of stochastic gradient and stochastic approximation algorithms. Assumption 2.2 is a noise condition. In this or a similar form, it is involved in most of the results on the convergence and convergence rate of stochastic gradient search and stochastic approximation. It holds for algorithms with Markovian dynamics (see the next section). It is also satisfied when $\{\xi_n\}_{n \geq 0}$ is a martingale-difference sequence. Assumption 2.3 is related to the stability of the gradient flow $d\theta/dt = -\nabla f(\theta)$, or more specifically, to the geometry of the set of stationary points S . In the area of differential geometry, relations (2.2) and (2.3) are known as the Lojasiewicz gradient inequality (see [18] and [19] for details). They hold if $f(\cdot)$ is analytic or subanalytic in an open vicinity of S (see [5], [19] for the proof; for the form of Lojasiewicz inequality appearing in Assumption 2.3 and (2.2) see [13, Theorem LI, page 775]; for the definition and properties of analytic and subanalytic functions, consult [5], [12]). Although analyticity and subanalyticity are fairly strong conditions, they hold for the objective functions of many stochastic gradient algorithms used in the areas of system identification, signal processing, machine learning, operations research and statistical inference. E.g., in this paper, we show that the objective functions associated with supervised learning and recursive prediction error identification are analytical (Sections 4 and 5). Moreover, in [28] (an extended version of this paper), we demonstrate the same property for temporal-difference learning algorithms. Furthermore, in [29], we show analyticity for the objective functions associated with recursive identification methods for hidden Markov models. It is also worth mentioning that the objective functions associated with recursive algorithms for principal and independent component analysis (as well as with many other adaptive signal processing algorithms) are

usually polynomial or rational, and hence, analytic, too (see e.g., [9] and references cited therein).

In order to state the main results of this section, we need further notation. For $\theta \in \mathbb{R}^{d_\theta}$, $C_\theta \in [1, \infty)$ stands for an upper bound of $\|\nabla f(\cdot)\|$ on $\{\theta' \in \mathbb{R}^{d_\theta} : \|\theta' - \theta\| \leq \delta_\theta\}$ and for a Lipschitz constant of $\nabla f(\cdot)$ on the same set. Moreover, p_θ , q_θ and r_θ are real numbers defines as

$$r_\theta = \begin{cases} 1/(2 - \mu_\theta), & \text{if } \mu_\theta < 2 \\ \infty, & \text{if } \mu_\theta = 2 \end{cases}, \quad p_\theta = \mu_\theta \min\{r, r_\theta\}, \quad q_\theta = \min\{r, r_\theta\} - 1 \quad (2.4)$$

(δ_θ , μ_θ are specified in Remark 2.1).

Our main results on the convergence and convergence rate of the recursion (2.1) are contained in the next two theorems.

THEOREM 2.1 (Convergence). *Let Assumptions 2.1 – 2.3 hold. Then, $\hat{\theta} = \lim_{n \rightarrow \infty} \theta_n$ exists and satisfies $\nabla f(\hat{\theta}) = 0$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$.*

THEOREM 2.2 (Convergence Rate). *Let Assumptions 2.1 – 2.3 hold. Then, there exists a random variable \hat{K} (which is a deterministic function of \hat{p} , $C_{\hat{\theta}}$, $M_{\hat{\theta}}$) such that $1 \leq \hat{K} < \infty$ everywhere and such that the following is true:*

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} \|\nabla f(\theta_n)\|^2 \leq \hat{K}(\varphi(\xi))^{\hat{\mu}}, \quad (2.5)$$

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} |f(\theta_n) - f(\hat{\theta})| \leq \hat{K}(\varphi(\xi))^{\hat{\mu}}, \quad (2.6)$$

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{q}} \|\theta_n - \hat{\theta}\| \leq \hat{K} \varphi(\xi) \quad (2.7)$$

w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$, where $\hat{\mu} = \mu_{\hat{\theta}}$, $\hat{p} = p_{\hat{\theta}}$, $\hat{q} = q_{\hat{\theta}}$, $\hat{r} = r_{\hat{\theta}}$ and

$$\varphi(\xi) = \begin{cases} \xi, & \text{if } r < \hat{r} \\ 1 + \xi, & \text{if } r = \hat{r} \\ 1, & \text{if } r > \hat{r} \end{cases}.$$

The proofs are provided in Section 6. As an immediate consequence of the previous theorems, we get the following corollaries:

COROLLARY 2.1. *Let Assumptions 2.1 – 2.3 hold. Then, the following is true:*

(i) $\|\nabla f(\theta_n)\|^2 = o(\gamma_n^{-\hat{p}})$, $|f(\theta_n) - f(\hat{\theta})| = o(\gamma_n^{-\hat{p}})$ and $\|\theta_n - \hat{\theta}\| = o(\gamma_n^{-\hat{q}})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\} \cap \{\xi = 0, \hat{r} > r\}$.

(ii) $\|\nabla f(\theta_n)\|^2 = O(\gamma_n^{-\hat{p}})$, $|f(\theta_n) - f(\hat{\theta})| = O(\gamma_n^{-\hat{p}})$ and $\|\theta_n - \hat{\theta}\| = O(\gamma_n^{-\hat{q}})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\} \cap \{\xi = 0, \hat{r} > r\}^c$.

(iii) $\|\nabla f(\theta_n)\|^2 = o(\gamma_n^{-p})$ and $|f(\theta_n) - f(\hat{\theta})| = o(\gamma_n^{-p})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$, where $p = \min\{1, r\}$.

In the literature on stochastic and deterministic optimization, the asymptotic behavior of gradient search is usually characterized by the convergence of sequences $\{\nabla f(\theta_n)\}_{n \geq 0}$, $\{f(\theta_n)\}_{n \geq 0}$ and $\{\theta_n\}_{n \geq 0}$ (see e.g., [3], [4], [23], [24] are references quoted therein). Similarly, the convergence rate can be described by the rates at which $\{\nabla f(\theta_n)\}_{n \geq 0}$, $\{f(\theta_n)\}_{n \geq 0}$ and $\{\theta_n\}_{n \geq 0}$ tend to the sets of their limit points. In the case of algorithm (2.1), this kind of information is provided by Theorems 2.1, 2.2 and Corollary 2.1. Theorem 2.1 claims that almost surely, algorithm (2.1) is point-convergence and does not exhibit limit cycles. Theorem 2.2 and Corollary 2.1 provide relatively tight upper bounds on the convergence rate of $\{\nabla f(\theta_n)\}_{n \geq 0}$, $\{f(\theta_n)\}_{n \geq 0}$

and $\{\theta_n\}_{n \geq 0}$. These bounds can be thought of as a combination of the convergence rate of the gradient flow $d\theta/dt = -\nabla f(\theta)$ (characterized by Lojasiewicz exponent μ_θ) and the rate of the noise averages $\sum_{i=n}^k \alpha_i \xi_i$ (expressed through parameter r and sequence $\{\gamma_n\}_{n \geq 0}$). Basically, Theorem 2.2 and Corollary 2.1 claim that the convergence rate of $\{\|\nabla f(\theta_n)\|^2\}_{n \geq 0}$ and $\{f(\theta_n)\}_{n \geq 0}$ is the slower of the rates $O(\gamma_n^{-\hat{r}\hat{\mu}})$ (the rate of the gradient flow $d\theta/dt = -\nabla f(\theta)$ sampled at instants $\{\gamma_n\}_{n \geq 0}$) and $O(\gamma_n^{-r\hat{\mu}})$ (the rate of the noise averages $\max_{k \geq n} \|\sum_{i=n}^k \alpha_i \xi_i\|^{\hat{\mu}}$).

Apparently, the results of Theorems 2.1, 2.2 and Corollary 2.1 are of a local nature: They hold only on the event where algorithm (2.1) is stable (i.e., where sequence $\{\theta_n\}_{n \geq 0}$ is bounded). Stating results on the convergence and convergence rate in such a local form is quite sensible due to the following reasons. The stability of stochastic gradient search is based on well-understood arguments which are rather different from the arguments used in the analysis of the convergence and convergence rate. Moreover and more importantly, it is straightforward to get a global version of the results provided in Theorems 2.1, 2.2 and Corollary 2.1 by combining the theorems with the methods used to verify or ensure the stability (e.g., with the results of [6] and [8]).

The point-convergence and convergence rate of stochastic gradient search (and stochastic approximation) have been the subject of a large number of papers and books (see [1], [14], [16], [24], [26] and references cited therein). Although the existing results provide a good insight into the asymptotic behavior and efficiency of stochastic gradient algorithms, they are based on fairly restrictive assumptions: Literally, they all require the objective function $f(\cdot)$ to have an isolated minimum θ_* (sometimes even to be strongly unimodal) such that Hessian $\nabla^2 f(\theta_*)$ is strictly positive definite and such that $\{\theta_n\}_{n \geq 0}$ visits the attraction domain of θ_* infinitely many times w.p.1. Unfortunately, in the case of high-dimensional high-nonlinear stochastic gradient algorithms (such as online machine learning and recursive identification), it is hard (if not impossible at all) to show even the existence of an isolated minimum, let alone the definiteness of $\nabla^2 f(\theta_*)$ and the infinitely often visits of $\{\theta_n\}_{n \geq 0}$ to the attraction domain of θ_* . Moreover and more importantly, these requirements are unlikely to be satisfied by a high-dimensional high-nonlinear algorithm, as the objective function associated with such an algorithm prones to manifolds of (non-isolated) minima and (non-isolated) saddles each of which is a potential limit point of the algorithm iterates (e.g., a recursive prediction error identification method exhibits this behavior when the candidate models are overparameterized or do not match the true system). Relying on the Lojasiewicz gradient inequality, Theorems 2.1, 2.2 and Corollary 2.1 overcome the described difficulties: Both theorems and their corollary allow the objective function $f(\cdot)$ to have multiple, non-isolated minima, impose no restriction on the values of $\nabla^2 f(\cdot)$ (notice that $\nabla^2 f(\cdot)$ cannot be strictly definite at a non-isolated minimum or maximum) and do not require (a priori) $\{\theta_n\}_{n \geq 0}$ to exhibit any particular behavior (i.e., to visit infinitely often the attraction domain of an isolated minimum). Moreover, they cover a broad class of complex stochastic gradient algorithms (see Sections 4 and 5; see also [28], [29]). To the best of our knowledge, these are the only results on the convergence and convergence rate of stochastic search which enjoy such features.

Regarding the results of Theorems 2.1, 2.2 and Corollary 2.1, it is worth mentioning that they are not just a combination of the Lojasiewicz inequality and the existing techniques for the asymptotic analysis of stochastic gradient search and stochastic approximation. On the contrary, the existing techniques seem to be completely in-

applicable to high-dimensional high-nonlinear stochastic gradient search. The reason comes out of the fact that these techniques crucially rely on the following Lyapunov function:

$$w(\theta) = (\theta - \theta_*)^T \nabla^2 f(\theta_*) (\theta - \theta_*),$$

where θ_* is an isolated minimum such that $\nabla^2 f(\theta_*)$ is strictly positive definite and such that the attraction domain of θ_* is visited by $\{\theta_n\}_{n \geq 0}$ infinitely many times w.p.1. In this paper, we take an entirely different approach whose main steps can be summarized as follows:

1. The convergence of $\{f(\theta_n)\}_{n \geq 0}$ is demonstrated.
2. A ‘singular’ Lyapunov function

$$v(\theta) = \begin{cases} (f(\theta) - \hat{f})^{-1/p}, & \text{if } f(\theta) > \hat{f} \\ 0, & \text{otherwise} \end{cases}$$

is constructed, where $\hat{f} = \lim_{n \rightarrow \infty} f(\theta_n)$ and p is a suitable positive constant. Relying on this function, the convergence rate of $\{f(\theta_n)\}_{n \geq 0}$ and $\{\nabla f(\theta_n)\}_{n \geq 0}$ is evaluated.

3. Using the results derived at Step 2, the convergence rate of $\sup_{k \geq n} \|\theta_k - \theta_n\|$ is assessed.

4. Applying the results of Step 3, the point-convergence of $\{\theta_n\}_{n \geq 0}$ is demonstrated. Then, refining the convergence rates derived at Steps 2 and 3, the results of Theorem 2.2 are obtained.

At the core of our approach is the singular Lyapunov function $v(\cdot)$. Although subtle techniques are needed to handle such a function (see Section 6), $v(\cdot)$ provides intuitively clear explanation of the results of Theorem 2.2 and Corollary 2.1. The explanation is based on the heuristic analysis of the following two cases.¹

CASE 2.1: $\liminf_{n \rightarrow \infty} \gamma_n^{\hat{\mu}r}(f(\theta_n)) - \hat{f} = -\infty$ and $\sup_{n \geq 0} \|\theta_n\| < \infty$, where $\hat{\mu}$ is defined in Theorem 2.2.

In this case, there exists an increasing integer sequence $\{n_k\}_{k \geq 0}$ such that $\lim_{k \rightarrow \infty} \gamma_{n_k}^{\hat{\mu}r}(f(\theta_{n_k})) - \hat{f} = -\infty$. Owing to Assumption 2.3, we have

$$\|\nabla f(\theta_n)\| \geq \left(|f(\theta_n) - \hat{f}| / \hat{M} \right)^{1/\hat{\mu}} \quad (2.8)$$

for sufficiently large n , where $\hat{M} = M_{\hat{\theta}}$. Consequently, $\lim_{k \rightarrow \infty} \gamma_{n_k}^r \|\nabla f(\theta_{n_k})\| = \infty$. On the other side, Taylor formula yields

$$\begin{aligned} f(\theta_n) &\approx f(\theta_{n_k}) - (\nabla f(\theta_{n_k}))^T \sum_{i=n_k}^{n-1} \alpha_i (\nabla f(\theta_i) + \xi_i) \\ &\approx f(\theta_{n_k}) - (\gamma_n - \gamma_{n_k}) \|\nabla f(\theta_{n_k})\|^2 - (\nabla f(\theta_{n_k}))^T \sum_{i=n_k}^{n-1} \alpha_i \xi_i \\ &\leq f(\theta_{n_k}) - \|\nabla f(\theta_{n_k})\| \left((\gamma_n - \gamma_{n_k}) \|\nabla f(\theta_{n_k})\| - \left\| \sum_{i=n_k}^{n-1} \alpha_i \xi_i \right\| \right) \end{aligned}$$

¹Throughout this analysis, we assume that Theorem 2.1 is true. We also assume $\sup_{k \geq n} \|\sum_{i=n}^k \alpha_i \xi_i\| = O(\gamma_n^{-r})$ when $n \rightarrow \infty$, which is slightly stronger than what Assumption 2.2 and Lemma 6.1 yield.

for $n \geq n_k$ and sufficiently large $k \geq 0$. Since $\gamma_n - \gamma_{n_k} \geq 1$ for $n > a(n_k, 1)$ and since

$$\sup_{k \geq n} \left\| \sum_{i=n}^k \alpha_i \xi_i \right\| = O(\gamma_n^{-r}) \quad (2.9)$$

when $n \rightarrow \infty$, we get $f(\theta_n) \leq f(\theta_{n_k}) < \hat{f}$ for $n > a(n_k, 1)$ and sufficiently large $k \geq 0$. However, this is not possible as $\lim_{n \rightarrow \infty} f(\theta_n) = \hat{f}$. Thus, Case 2.1 cannot occur.

CASE 2.2: $\limsup_{n \rightarrow \infty} \gamma_n^p (f(\theta_n) - \hat{f}) = \infty$, $p < \hat{\mu} \min\{r, \hat{r}\}$ and $\sup_{n \geq 0} \|\theta_n\| < \infty$, where $\hat{\mu}, \hat{r}$ are defined in Theorem 2.2.

Similarly as in the previous case, there exists an increasing integer sequence $\{n_k\}_{k \geq 0}$ such that $\lim_{k \rightarrow \infty} \gamma_{n_k}^p (f(\theta_{n_k}) - \hat{f}) = \infty$. Then, (2.8) implies

$$\lim_{k \rightarrow \infty} \gamma_{n_k}^r (f(\theta_{n_k}) - \hat{f}) \geq \lim_{k \rightarrow \infty} \gamma_{n_k}^{p/\hat{\mu}} (f(\theta_{n_k}) - \hat{f}) = \infty \quad (2.10)$$

(notice that $p/\hat{\mu} \leq r$). On the other side, Taylor formula and (2.8) yield

$$\begin{aligned} v(\theta_n) &\approx v(\theta_{n_k}) + \frac{(\nabla f(\theta_{n_k}))^T}{p(f(\theta_{n_k}) - \hat{f})^{1+1/p}} \sum_{i=n_k}^{n-1} \alpha_i (\nabla f(\theta_i) + \xi_i) \\ &\approx v(\theta_{n_k}) + \frac{1}{p(f(\theta_{n_k}) - \hat{f})^{1+1/p}} \left((\gamma_n - \gamma_{n_k}) \|\nabla f(\theta_{n_k})\|^2 + (f(\theta_{n_k}))^T \sum_{i=n_k}^{n-1} \alpha_i \xi_i \right) \\ &\geq v(\theta_{n_k}) + \frac{\gamma_n - \gamma_{n_k}}{2p\hat{M}^{2/\hat{\mu}}(f(\theta_{n_k}) - \hat{f})^{1+1/p-2/\hat{\mu}}} \\ &\quad + \frac{\|\nabla f(\theta_{n_k})\|}{p(f(\theta_{n_k}) - \hat{f})^{1+1/p}} \left(\frac{(\gamma_n - \gamma_{n_k}) \|\nabla f(\theta_{n_k})\|}{2} - \left\| \sum_{i=n_k}^{n-1} \alpha_i \xi_i \right\| \right) \end{aligned} \quad (2.11)$$

for $n \geq n_k$ and sufficiently large $k \geq 0$. Since $\lim_{n \rightarrow \infty} f(\theta_n) = \hat{f}$ and since

$$1 + 1/p - 2/\hat{\mu} = 1/p - \hat{\mu}/\hat{r} \geq 0,$$

relations (2.9) – (2.11) imply

$$v(\theta_n) \geq v(\theta_{n_k}) + \hat{N}(\gamma_n - \gamma_{n_k})$$

for $n > a(n_k, 1)$, sufficiently large $k \geq 0$ and $\hat{N} = 1/(2p\hat{M}^{2/\hat{\mu}})$. Consequently,

$$f(\theta_n) - \hat{f} \leq \left(v(\theta_{n_k}) + \hat{N}(\gamma_n - \gamma_{n_k}) \right)^{-p} \quad (2.12)$$

for $n > a(n_k, 1)$ and sufficiently large $k \geq 0$. However, this is impossible, as (2.12) yields $\limsup_{n \rightarrow \infty} \gamma_n^p (f(\theta_n) - \hat{f}) < \infty$. Hence, Case 2.2 cannot happen.

As none of Cases 2.1 and 2.2 is possible, we conclude that $f(\theta_n)$ converges to \hat{f} at the rate $O(\gamma_n^{-\hat{p}})$. Since $\gamma_k - \gamma_n \geq 1$ for $k > a(n, 1)$ and since

$$\begin{aligned} f(\theta_k) - f(\theta_n) &\approx -(\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 - (\nabla f(\theta_n))^T \sum_{i=n}^{k-1} \alpha_i \xi_i \\ &\leq -((\gamma_k - \gamma_n) - 1/2) \|\nabla f(\theta_n)\|^2 + \frac{1}{2} \left\| \sum_{i=n}^{k-1} \alpha_i \xi_i \right\|^2 \end{aligned}$$

for $k \geq n$ and sufficiently large $n \geq 0$, we deduce

$$\|\nabla f(\theta_n)\|^2 \leq -2(f(\theta_k) - f(\theta_n)) + \left\| \sum_{i=n}^{k-1} \alpha_i \xi_i \right\|^2$$

for $k > a(n, 1)$ and sufficiently large $n \geq 0$. As an immediate consequence, we have that $\|\nabla f(\theta_n)\|^2$ converges to zero at the rate $O(\gamma_n^{-\hat{p}})$. The evaluation of the convergence rate of $\{\theta_n\}_{n \geq 0}$ is much more complicated (so that it cannot briefly be summarized here — the details are provided in Lemmas 6.6, 6.10) and is based on the following reasoning:

$$\begin{aligned} \|\theta_k - \theta_n\| &\leq \left\| \theta_k - \theta_n + \sum_{i=n}^{k-1} \alpha_i \xi_i \right\| + \left\| \sum_{i=n}^{k-1} \alpha_i \xi_i \right\| \\ &= \left\| \sum_{i=n}^{k-1} \alpha_i \nabla f(\theta_i) \right\| + \left\| \sum_{i=n}^{k-1} \alpha_i \xi_i \right\| \\ &\approx (\gamma_k - \gamma_n) \|\nabla f(\theta_n)\| + \left\| \sum_{i=n}^{k-1} \alpha_i \xi_i \right\| \\ &\approx -\frac{1}{\|\nabla f(\theta_n)\|} \left(f(\theta_k) - f(\theta_n) + (\nabla f(\theta_n))^T \sum_{i=n}^{k-1} \alpha_i \xi_i \right) + \left\| \sum_{i=n}^{k-1} \alpha_i \xi_i \right\| \\ &\leq -\frac{f(\theta_k) - f(\theta_n)}{\|\nabla f(\theta_n)\|} + 2 \left\| \sum_{i=n}^{k-1} \alpha_i \xi_i \right\| \end{aligned}$$

where $k \geq n$ and $n \geq 0$ is sufficiently large.

The heuristic analysis of Cases 2.1 and 2.2 carried out above indicates that the convergence rates of $\{f(\theta_n)\}_{n \geq 0}$ and $\{\nabla f(\theta_n)\}_{n \geq 0}$ reported in Theorem 2.2 are rather tight (if not optimal; for the discussion on the tightness of the rate of $\{\theta_n\}_{n \geq 0}$, see Remark 6.4). The same conclusion is suggested by the following two special cases:

CASE 2.3: $\xi_n = 0$ for each $n \geq 0$.

Due to Assumption 2.3 and (2.8), we have

$$\frac{d(f(\theta(t)) - \hat{f})}{dt} = -\|\nabla f(\theta(t))\|^2 \leq -\left(\frac{f(\theta(t)) - \hat{f}}{\hat{M}} \right)^{2/\hat{\mu}}$$

for a solution $\theta(\cdot)$ of $d\theta/dt = -\nabla f(\theta)$ satisfying $\lim_{t \rightarrow \infty} f(\theta(t)) = \hat{f}$ and $\theta([0, \infty)) \subseteq \{\theta \in \mathbb{R}^{d_\theta} : \|\theta - \hat{\theta}\| \leq \delta_{\hat{\theta}}\}$. Consequently,

$$f(\theta(t)) - \hat{f} = O(t^{-\hat{\mu}/(2-\hat{\mu})}) = O(t^{-\hat{\mu}\hat{r}}).$$

As $\{\theta_n\}_{n \geq 0}$ is asymptotically equivalent to $\theta(\cdot)$ sampled at instances $\{\gamma_n\}_{n \geq 0}$, we get $f(\theta_n) - \hat{f} = O(\gamma_n^{-\hat{\mu}\hat{r}})$. The same result is implied by Theorem 2.1 and Corollary 2.1.

CASE 2.4: $f(\theta) = \theta^T A \theta$, where A is a strictly positive definite matrix.

Recursion (2.1) reduces to a linear stochastic approximation algorithm in this case. For such an algorithm, the tightest bound on the convergence rate of $\{f(\theta_n)\}_{n \geq 0}$ and $\{\|\nabla f(\theta_n)\|^2\}_{n \geq 0}$ is $O(\gamma_n^{-2r})$ if $\xi > 0$ and $o(\gamma_n^{-2r})$ if $\xi = 0$ (see [27]). The same rate is predicted by Theorem 2.2 and Corollary 2.1.

3. Stochastic Gradient Algorithms with Markovian Dynamics. In order to illustrate the results of Section 2 and to set up a framework for the analysis carried out in Sections 4 and 5, we apply Theorems 2.1, 2.2 and Corollary 2.1 to stochastic gradient algorithms with Markovian dynamics. These algorithms are defined by the following difference equation:

$$\theta_{n+1} = \theta_n - \alpha_n F(\theta_n, Z_{n+1}), \quad n \geq 0. \quad (3.1)$$

In this recursion, $F : \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_z} \rightarrow \mathbb{R}^{d_\theta}$ is a Borel-measurable function, while $\{\alpha_n\}_{n \geq 0}$ is a sequence of positive real numbers. θ_0 is an \mathbb{R}^{d_θ} -valued random variable defined on a probability space (Ω, \mathcal{F}, P) , while $\{Z_n\}_{n \geq 0}$ is an \mathbb{R}^{d_z} -valued stochastic process defined on the same probability space. $\{Z_n\}_{n \geq 0}$ is a Markov process controlled by $\{\theta_n\}_{n \geq 0}$, i.e., there exists a family of transition probability kernels $\{\Pi_\theta(\cdot, \cdot)\}_{\theta \in \mathbb{R}^{d_\theta}}$ (defined on \mathbb{R}^{d_z}) such that

$$P(Z_{n+1} \in B | \theta_0, Z_0, \dots, \theta_n, Z_n) = \Pi_{\theta_n}(Z_n, B)$$

w.p.1 for any Borel-measurable set $B \subseteq \mathbb{R}^{d_z}$ and $n \geq 0$. In the context of stochastic gradient search, $F(\theta_n, Z_{n+1})$ is regarded to as an estimator of $\nabla f(\theta_n)$.

The algorithm (3.1) is analyzed under the following assumptions.

ASSUMPTION 3.1. $\lim_{n \rightarrow \infty} \alpha_n = 0$, $\limsup_{n \rightarrow \infty} |\alpha_{n+1}^{-1} - \alpha_n^{-1}| < \infty$ and $\sum_{n=0}^{\infty} \alpha_n = \infty$. There exists a real number $r \in (1, \infty)$ such that $\sum_{n=0}^{\infty} \alpha_n^2 \gamma_n^{2r} < \infty$.

ASSUMPTION 3.2. There exist a differentiable function $f : \mathbb{R}^{d_\theta} \rightarrow \mathbb{R}$ and a Borel-measurable function $\tilde{F} : \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_z} \rightarrow \mathbb{R}^{d_\theta}$ such that $\nabla f(\cdot)$ is locally Lipschitz continuous and such that

$$F(\theta, z) - \nabla f(\theta) = \tilde{F}(\theta, z) - (\Pi \tilde{F})(\theta, z)$$

for each $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$, where $(\Pi \tilde{F})(\theta, z) = \int \tilde{F}(\theta, z') \Pi_\theta(z, dz')$.

ASSUMPTION 3.3. For any compact set $Q \subset \mathbb{R}^{d_\theta}$ and $s \in (0, 1)$, there exists a Borel-measurable function $\varphi_{Q,s} : \mathbb{R}^{d_z} \rightarrow [1, \infty)$ such that

$$\begin{aligned} \max\{\|F(\theta, z)\|, \|\tilde{F}(\theta, z)\|, \|(\Pi \tilde{F})(\theta, z)\|\} &\leq \varphi_{Q,s}(z), \\ \|(\Pi \tilde{F})(\theta', z) - (\Pi \tilde{F})(\theta'', z)\| &\leq \varphi_{Q,s}(z) \|\theta' - \theta''\|^s \end{aligned}$$

for all $\theta, \theta', \theta'' \in Q$, $z \in \mathbb{R}^{d_z}$.

ASSUMPTION 3.4. Given a compact set $Q \subset \mathbb{R}^{d_\theta}$ and $s \in (0, 1)$,

$$\sup_{n \geq 0} E(\varphi_{Q,s}^2(Z_n) I_{\{\tau_Q \geq n\}} | \theta_0 = \theta, Z_0 = z) < \infty$$

for all $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$, where $\tau_Q = \inf\{n \geq 0 : \theta_n \notin Q\}$.

The main results on the convergence rate of recursion (3.1) are contained in the next theorem.

THEOREM 3.1. Let Assumptions 3.1 – 3.4 hold, and suppose that $f(\cdot)$ (introduced in Assumption 3.2) satisfies Assumption 2.3. Then, the following is true:

- (i) $\hat{\theta} = \lim_{n \rightarrow \infty} \theta_n$ exists and satisfies $\nabla f(\hat{\theta}) = 0$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$.
- (ii) $\|\nabla f(\theta_n)\|^2 = o(\gamma_n^{-\hat{p}})$, $|f(\theta_n) - f(\hat{\theta})| = o(\gamma_n^{-\hat{p}})$ and $\|\theta_n - \hat{\theta}\| = o(\gamma_n^{-\hat{q}})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\} \cap \{\hat{r} > r\}$.
- (iii) $\|\nabla f(\theta_n)\|^2 = O(\gamma_n^{-\hat{p}})$, $|f(\theta_n) - f(\hat{\theta})| = O(\gamma_n^{-\hat{p}})$ and $\|\theta_n - \hat{\theta}\| = O(\gamma_n^{-\hat{q}})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\} \cap \{\hat{r} \leq r\}$.

(iv) $\|\nabla f(\theta_n)\|^2 = o(\gamma_n^{-p})$ and $|f(\theta_n) - f(\hat{\theta})| = o(\gamma_n^{-p})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$.

The proof is provided in Section 7. p , \hat{p} , \hat{q} and \hat{r} are defined in Theorem 2.2 and Corollary 2.1.

Assumption 3.1 is related to the sequence $\{\alpha_n\}_{n \geq 0}$. It holds if $\alpha_n = 1/n^a$ for $n \geq 1$, where $a \in (3/4, 1]$ is a constant (in that case, $\gamma_n = O(n^{1-a})$ for $n \rightarrow \infty$, while r can be any number satisfying $0 < r < (a - 1/2)/(1 - a)$). On the other side, Assumptions 3.2 – 3.4 correspond to the stochastic process $\{Z_n\}_{n \geq 0}$ and are quite standard for the asymptotic analysis of stochastic approximation algorithms with Markovian dynamics. Assumptions 3.2 – 3.4 have been introduced by Metivier and Priouret in [20] (see also [1, Part II]), and later generalized by Kushner and his co-workers (see [14] and references cited therein). However, neither the results of Metivier and Priouret, nor the results of Kushner and his co-workers provide any information on the point-convergence and convergence rate of stochastic gradient search in the case of multiple, non-isolated minima.

Regarding Theorem 3.1, the following note is also in order. As already mentioned in the beginning of the section, the purpose of the theorem is illustrating the results of Theorem 2.1 and providing a framework for studying the examples presented in the next sections. Since these examples perfectly fit into the framework developed by Metivier and Priouret, more general assumptions and settings of [14] are not considered here in order just to keep the exposition as concise as possible.

4. Example 1: Supervised Learning. In this section, online algorithms for supervised learning in feedforward neural networks are analyzed using the results of Theorems 2.2 and 3.1. To avoid unnecessary technical details and complicated notation, only two-layer perceptrons are considered here. However, the obtained results can be extended to other feedforward neural networks such radial basis function networks.

The input-output function of a two-layer perceptron can be defined as

$$G_\theta(x) = \sum_{i=1}^M a_i \psi \left(\sum_{j=1}^N b_{i,j} x_j \right).$$

Here, $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is a differentiable function, while M and N are positive integers. a_1, \dots, a_M , $b_{1,1}, \dots, b_{M,N}$ and x_1, \dots, x_N are real numbers, while $\theta = [a_1 \dots a_M \ b_{1,1} \dots b_{M,N}]^T$, $x = [x_1 \dots x_N]^T$ and $d_\theta = M(N + 1)$. $\psi(\cdot)$ represents the network activation function. x is the network input, while $G_\theta(x)$ is the output. θ is the vector of the network parameters to be tuned through the process of supervised learning.

Let $\pi(\cdot, \cdot)$ be a probability measure on $\mathbb{R}^N \times \mathbb{R}$, while

$$f(\theta) = \frac{1}{2} \int (y - G_\theta(x))^2 \pi(dx, dy)$$

for $\theta \in \mathbb{R}^{d_\theta}$. Then, the mean-square error based supervised learning in feedforward neural networks can be described as the minimization of $f(\cdot)$ in a situation when only samples from $\pi(\cdot, \cdot)$ are available. For more details on neural networks and supervised learning, see e.g., [10], [11] and references cited therein.

Function $f(\cdot)$ is usually minimized by the following stochastic gradient algorithm:

$$\theta_{n+1} = \theta_n + \alpha_n (Y_n - G_{\theta_n}(X_n)) H_{\theta_n}(X_n), \quad n \geq 0. \quad (4.1)$$

In this recursion, $\{\alpha_n\}_{n \geq 0}$ is a sequence of positive real numbers, while $H_\theta(\cdot) = \nabla_\theta G_\theta(\cdot)$. θ_0 is an \mathbb{R}^{d_θ} -valued random variable defined on a probability space (Ω, \mathcal{F}, P) , while $\{(X_n, Y_n)\}_{n \geq 0}$ is an $\mathbb{R}^N \times \mathbb{R}$ -valued stochastic process defined on the same probability space. In the context of supervised learning, $\{(X_n, Y_n)\}_{n \geq 0}$ is regarded to as a training sequence.

The asymptotic behavior of algorithm (4.1) is analyzed under the following assumptions:

ASSUMPTION 4.1. $\psi(\cdot)$ is real-analytic. Moreover, $\psi(\cdot)$ has a (complex-valued) continuation $\hat{\psi}(\cdot)$ with the following properties:

- (i) $\hat{\psi}(z)$ maps $z \in \mathbb{C}$ into \mathbb{C} (\mathbb{C} denotes the set of complex numbers).
- (ii) $\hat{\psi}(x) = \psi(x)$ for all $x \in \mathbb{R}$.
- (iii) There exist real numbers $\varepsilon \in (0, 1)$, $K \in [1, \infty)$ such that $\hat{\psi}(\cdot)$ is analytic on $\hat{V}_\varepsilon = \{z \in \mathbb{C} : d(z, \mathbb{R}) \leq \varepsilon\}$, and such that

$$\max\{|\hat{\psi}(z)|, |\hat{\psi}'(z)|\} \leq K$$

for all $z \in \hat{V}_\varepsilon$ ($\hat{\psi}'(\cdot)$ is the first derivative of $\hat{\psi}(\cdot)$).

ASSUMPTION 4.2. $\{(X_n, Y_n)\}_{n \geq 0}$ are i.i.d. random variables distributed according the probability measure $\pi(\cdot, \cdot)$. There exists a real number $L \in [1, \infty)$ such that $\|X_0\| \leq L$ and $|Y_0| \leq L$ w.p.1.

Our main results on the properties of objective function $f(\cdot)$ and algorithm (4.1) are contained in the next two theorems.

THEOREM 4.1. Let Assumptions 4.1 and 4.2 hold. Then, $f(\cdot)$ is analytic on entire \mathbb{R}^{d_θ} , i.e., it satisfies Assumption 2.3.

THEOREM 4.2. Let Assumptions 3.1, 4.1 and 4.2 hold. Then, the following is true:

- (i) $\hat{\theta} = \lim_{n \rightarrow \infty} \theta_n$ exists and satisfies $\nabla f(\hat{\theta}) = 0$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$.
- (ii) $\|\nabla f(\theta_n)\|^2 = o(\gamma_n^{-\hat{p}})$, $|f(\theta_n) - f(\hat{\theta})| = o(\gamma_n^{-\hat{p}})$ and $\|\theta_n - \hat{\theta}\| = o(\gamma_n^{-\hat{q}})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\} \cap \{\hat{r} > r\}$.
- (iii) $\|\nabla f(\theta_n)\|^2 = O(\gamma_n^{-\hat{p}})$, $|f(\theta_n) - f(\hat{\theta})| = O(\gamma_n^{-\hat{p}})$ and $\|\theta_n - \hat{\theta}\| = O(\gamma_n^{-\hat{q}})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\} \cap \{\hat{r} \leq r\}$.
- (iv) $\|\nabla f(\theta_n)\|^2 = o(\gamma_n^{-p})$ and $|f(\theta_n) - f(\hat{\theta})| = o(\gamma_n^{-p})$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$.

The proofs are provided in Section 8. p , \hat{p} , \hat{q} and \hat{r} are defined in Theorem 2.2 and Corollary 2.1.

Assumption 4.1 is related to the network activation function. It holds when $\psi(\cdot)$ is a logistic function² or a standard Gaussian density³, which are the most popular activation functions in feedforward neural networks. Assumption 4.2 corresponds to the training sequence $\{(X_n, Y_n)\}_{n \geq 0}$, and is common for the analysis of supervised learning.

²Complex-valued logistic function can be defined as $h(z) = (1 + \exp(-z))^{-1}$ for $z \in \mathbb{C}$. Since

$$|1 + \exp(-z)|^2 = 1 + \exp(-2\operatorname{Re}(z)) + 2\exp(-\operatorname{Re}(z))\cos(\operatorname{Im}(z)) \geq 1 + \exp(-2\operatorname{Re}(z))$$

when $|\operatorname{Im}(z)| \leq \pi/2$, $h(\cdot)$ is analytical on $\{z \in \mathbb{C} : d(z, \mathbb{R}) \leq \pi/2\}$. Due to the same reason, $\max\{|h(z)|, |h'(z)|\} \leq 1$ on $\{z \in \mathbb{C} : d(z, \mathbb{R}) \leq \pi/2\}$.

³Complex-valued standard Gaussian density is defined by $h(z) = (2\pi)^{-1/2} \exp(-z^2/2)$ for $z \in \mathbb{C}$. It is analytical on entire \mathbb{C} . As

$$(1 + |z|) \exp(-z^2/2) \leq (1 + |\operatorname{Re}(z)| + |\operatorname{Im}(z)|) \exp(-\operatorname{Re}^2(z)/2 + \operatorname{Im}^2(z)/2) \leq 3e$$

when $|\operatorname{Im}(z)| \leq 1$, we have $\max\{|h(z)|, |h'(z)|\} \leq 3e$ on $\{z \in \mathbb{C} : d(z, \mathbb{R}) \leq 1\}$.

The asymptotic properties of supervised learning algorithms have been studied in a large number of papers (see [10], [11] and references cited therein). Unfortunately, the available literature does not provide any information on the point-convergence and convergence rate which can be verified for feedforward neural networks with nonlinear activation functions. The main difficulty comes out of the fact that the existing results on the convergence and convergence rate of stochastic gradient search require the objective function $f(\cdot)$ to have an isolated minimum θ_* such that $\nabla^2 f(\theta_*)$ is strictly positive definite and such that $\{\theta_n\}_{n \geq 0}$ visits the attraction domain of θ_* infinitely many times w.p.1. Since $f(\cdot)$ is highly nonlinear, these requirements are not only hard (if possible at all) to show, but are rather likely not to hold. Theorem 4.2 does not invoke any of such requirements and covers some of the most widely used feedforward neural networks.

5. Example 2: Identification of Linear Stochastic Dynamical Systems.

In this section, the general results presented in Sections 2 and 3 are applied to the asymptotic analysis of recursive prediction error algorithms for identification of linear stochastic systems. To avoid unnecessary technical details and complicated notation, only the identification of one dimensional ARMA models is considered here. However, it is straightforward to generalize the obtained results to any linear stochastic system.

To state the problem of the recursive prediction error identification in ARMA models, we use the following notation. M and N are positive integers. For $a_1, \dots, a_M \in \mathbb{R}$ and $b_1, \dots, b_N \in \mathbb{R}$, let

$$A_\theta(z) = 1 - \sum_{k=1}^M a_k z^{-k}, \quad B_\theta(z) = 1 + \sum_{k=1}^N b_k z^{-k},$$

where $\theta = [a_1 \dots a_M \ b_1 \dots b_N]^T$ and $z \in \mathbb{C}$ (\mathbb{C} denotes the set of complex numbers). Moreover, let $d_\theta = M + N$ and

$$\Theta = \{\theta \in \mathbb{R}^{d_\theta} : B_\theta(z) = 0 \Rightarrow |z| > 1\}.$$

$\{Y_n\}_{n \geq 0}$ is a real-valued signal generated by the actual system (i.e., by the system being identified). For $\theta \in \Theta$, $\{Y_n^\theta\}_{n \geq 0}$ is the output of the ARMA model

$$A_\theta(q)Y_n^\theta = B_\theta(q)W_n, \quad n \geq 0, \quad (5.1)$$

where $\{W_n\}_{n \geq 0}$ is a real-valued white noise and q^{-1} is the backward time-shift operator. $\{\varepsilon_n^\theta\}_{n \geq 0}$ is the process generated by the recursion

$$B_\theta(q)\varepsilon_n^\theta = A_\theta(q)Y_n, \quad n \geq 0, \quad (5.2)$$

while $\hat{Y}_n^\theta = Y_n - \varepsilon_n^\theta$ for $n \geq 0$. \hat{Y}_n^θ represents a mean-square optimal estimate of Y_n given Y_0, \dots, Y_{n-1} (which the model (5.1) can provide; for details see e.g., [16], [17]). Consequently, ε_n^θ can be interpreted as the estimation error of \hat{Y}_n^θ .

The parametric identification in ARMA models can be stated as follows: Given a realization of $\{Y_n\}_{n \geq 0}$, estimate the values of θ for which the model (5.1) provides the best approximation to the signal $\{Y_n\}_{n \geq 0}$. If the identification is based on the prediction error principle, this estimation problem reduces to the minimization of the asymptotic mean-square prediction error

$$f(\theta) = \frac{1}{2} \lim_{n \rightarrow \infty} E((\varepsilon_n^\theta)^2)$$

over Θ . As the asymptotic value of the second moment of ε_n^θ is rarely available analytically, $f(\cdot)$ is minimized by a stochastic gradient (or stochastic Newton) algorithm. Such an algorithm is defined by the following difference equations:

$$\phi_n = [Y_n \cdots Y_{n-M+1} \varepsilon_n \cdots \varepsilon_{n-N+1}]^T, \quad (5.3)$$

$$\varepsilon_{n+1} = Y_{n+1} - \phi_n^T \theta_n, \quad (5.4)$$

$$\psi_{n+1} = \phi_n - [\psi_n \cdots \psi_{n-N+1}]^T D \theta_n, \quad (5.5)$$

$$\theta_{n+1} = \theta_n + \alpha_n \psi_{n+1} \varepsilon_{n+1}, \quad n \geq 0. \quad (5.6)$$

In this recursion, $\{\alpha_n\}_{n \geq 0}$ denotes a sequence of positive reals. D is an $N \times (M+N)$ matrix whose entries are $d_{i,j} = 1$ if $j = M+i$, $1 \leq i \leq N$ and $d_{i,j} = 0$ otherwise. $\{Y_n\}_{n \geq -M}$ is a real-valued stochastic process defined on a probability space (Ω, \mathcal{F}, P) , while $\theta_0 \in \Theta$, $\varepsilon_0, \dots, \varepsilon_{1-N} \in \mathbb{R}$ and $\psi_0, \dots, \psi_{1-N} \in \mathbb{R}^{d_\theta}$ are random variables defined on the same probability space. $\theta_0, \varepsilon_0, \dots, \varepsilon_{1-N}, \psi_0, \dots, \psi_{1-N}$ represent the initial conditions of the algorithm (5.3) – (5.6).

In the literature on system identification, recursion (5.3) – (5.6) is known as the recursive prediction error algorithm for ARMA models (for more details see [16], [17] and references cited therein). It usually involves a projection (or truncation) device which ensures that estimates $\{\theta_n\}_{n \geq 0}$ remain in Θ . However, in order to avoid unnecessary technical details and to keep the exposition as concise as possible, this aspect of algorithm (5.3) – (5.6) is not discussed here. Instead, similarly as in [15] – [17], we state our asymptotic results (Theorem 5.2) in a local form.

Algorithm (5.3) – (5.6) is analyzed under the following assumptions:

ASSUMPTION 5.1. *There exist a positive integer L , a matrix $A \in \mathbb{R}^{L \times L}$, a vector $b \in \mathbb{R}^L$ and \mathbb{R}^L -valued stochastic processes $\{X_n\}_{n > -M}$, $\{V_n\}_{n > -M}$ (defined on (Ω, \mathcal{F}, P)) such that the following holds:*

- (i) $X_{n+1} = AX_n + V_n$ and $Y_n = b^T X_n$ for $n > -M$.
- (ii) The eigenvalues of A lie in $\{z \in \mathbb{C} : |z| < 1\}$.
- (iii) $\{V_n\}_{n \geq -M}$ are i.i.d. and independent of $\theta_0, X_{1-M}, \varepsilon_0, \dots, \varepsilon_{1-N}, \psi_0, \dots, \psi_{1-N}$.
- (iv) $E\|V_0\|^4 < \infty$.

ASSUMPTION 5.2. *For any compact set $Q \subset \Theta$,*

$$\sup_{n \geq 0} E((\varepsilon_n^4 + \|\psi_n\|^4) I_{\{\tau_Q \geq n\}}) < \infty, \quad (5.7)$$

where $\tau_Q = \inf\{n \geq 0 : \theta_n \notin Q\}$.

Our main result on the analyticity of $f(\cdot)$ is contained in the next theorem.

THEOREM 5.1. *Suppose that $\{Y_n\}_{n \geq 0}$ is a weakly stationary process such that*

$$\sum_{n=0}^{\infty} |\text{Cov}(Y_0, Y_n)| < \infty.$$

Then, $f(\cdot)$ is analytic on entire Θ , i.e., the following is true: For any compact set $Q \subset \Theta$ and any $a \in f(Q)$, there exist real numbers $\delta_{Q,a} \in (0, 1]$, $\mu_{Q,a} \in (1, 2]$, $M_{Q,a} \in [1, \infty)$ such that (2.2) is satisfied for all $\theta \in Q$ fulfilling $|f(\theta) - a| \leq \delta_{Q,a}$.

Let Λ is the event defined by

$$\Lambda = \left\{ \sup_{n \geq 0} \|\theta_n\| < \infty, \inf_{n \geq 0} d(\theta_n, \partial\Theta) > 0 \right\}.$$

Then, our main result on the convergence and convergence rate of algorithm (5.3) – (5.6) reads as follows.

THEOREM 5.2. *Let Assumptions 3.1, 5.1 and 5.2 hold. Then, the following is true:*

- (i) $\hat{\theta} = \lim_{n \rightarrow \infty} \theta_n$ exists and satisfies $\nabla f(\hat{\theta}) = 0$ w.p.1 on Λ .
- (ii) $\|\nabla f(\theta_n)\|^2 = o(\gamma_n^{-\hat{p}})$, $|f(\theta_n) - f(\hat{\theta})| = o(\gamma_n^{-\hat{p}})$ and $\|\theta_n - \hat{\theta}\| = o(\gamma_n^{-\hat{q}})$ w.p.1 on $\Lambda \cap \{\hat{r} > r\}$.
- (iii) $\|\nabla f(\theta_n)\|^2 = O(\gamma_n^{-\hat{p}})$, $|f(\theta_n) - f(\hat{\theta})| = O(\gamma_n^{-\hat{p}})$ and $\|\theta_n - \hat{\theta}\| = O(\gamma_n^{-\hat{q}})$ w.p.1 on $\Lambda \cap \{\hat{r} \leq r\}$.
- (iv) $\|\nabla f(\theta_n)\|^2 = o(\gamma_n^{-p})$ and $|f(\theta_n) - f(\hat{\theta})| = o(\gamma_n^{-p})$ w.p.1 on Λ .

The proofs are provided in Section 9. p , \hat{p} , \hat{q} and \hat{r} are defined in Theorem 2.2 and Corollary 2.1.

Assumption 5.1 corresponds to the signal $\{Y_n\}_{n \geq 0}$. It is quite common for the asymptotic analysis of recursive identification algorithm (see e.g., [1, Part I]) and cover all stable linear Markov models. Assumption 5.2 is related to the stability of subrecursion (5.3) – (5.5) and its output $\{\varepsilon_n\}_{n \geq 0}$, $\{\psi_n\}_{n \geq 0}$. In this or a similar form, Assumption 5.2 is involved in most of the asymptotic results on the recursive prediction error identification algorithms. E.g., [16, Theorems 4.1 – 4.3] (which are probably the most general results of this kind) require sequence $\{(\varepsilon_n, \psi_n)\}_{n \geq 0}$ to visit a fixed compact set infinitely often w.p.1 on event Λ . When $\{Y_n\}_{n \geq 0}$ is generated by a stable linear Markov system, such a requirement is practically equivalent to (5.7).

Various aspects of recursive prediction error identification in linear stochastic systems have been the subject of numerous papers and books (see [16], [17] and references cited therein). Despite providing a deep insight into the asymptotic behavior of recursive prediction error identification algorithms, the available results do not offer information about the point-convergence and convergence rate which can be verified for models of a moderate or high order (e.g., M and N are three or above). The main difficulty is the same as in the case of supervised learning. The existing results on the convergence and convergence rate of stochastic gradient search require $f(\cdot)$ to have an isolated minimum θ_* such that $\nabla^2 f(\theta_*)$ is strictly positive definite and such that $\{\theta_n\}_{n \geq 0}$ visits the attraction domain of θ_* infinitely many times w.p.1. Unfortunately, $f(\cdot)$ is so complex (even for relatively small M and N) that these requirements are not only impossible to verify, but are likely not to be true. Apparently, Theorem 5.2 relies on none of them.

Regarding Theorems 5.1 and 5.2, it should be mentioned that these results can be generalized in several ways. E.g., it is straightforward to extend them to practically any stable multiple-input, multiple-output linear system. Moreover, it is possible to show that the results also hold for signals $\{Y_n\}_{n \geq 0}$ satisfying mixing conditions of the type [16, Condition S1, p. 169].

6. Proof of Theorems 2.1 and 2.2. In this section, the following notation is used. Let Λ be the event

$$\Lambda = \left\{ \sup_{n \geq 0} \|\theta_n\| < \infty \right\}.$$

For $\varepsilon \in (0, \infty)$, let

$$\varphi_\varepsilon(\xi) = \varphi(\xi) + \varepsilon.$$

For $0 \leq n < k$, let $\zeta_{n,n} = \zeta'_{n,n} = \zeta''_{n,n} = 0$, $\phi_{n,n} = \phi'_{n,n} = \phi''_{n,n} = 0$ and

$$\begin{aligned}\zeta'_{n,k} &= \sum_{i=n}^{k-1} \alpha_i \xi_i, \\ \zeta''_{n,k} &= \sum_{i=n}^{k-1} \alpha_i (\nabla f(\theta_i) - \nabla f(\theta_n)), \\ \zeta_{n,k} &= \zeta'_{n,k} + \zeta''_{n,k}, \\ \phi'_{n,k} &= (\nabla f(\theta_n))^T \zeta_{n,k}, \\ \phi''_{n,k} &= - \int_0^1 (\nabla f(\theta_n + s(\theta_k - \theta_n)) - \nabla f(\theta_n))^T (\theta_k - \theta_n) ds, \\ \phi_{n,k} &= \phi'_{n,k} + \phi''_{n,k}.\end{aligned}$$

Then, it is straightforward to show

$$\begin{aligned}\theta_k - \theta_n &= - \sum_{i=n}^{k-1} \alpha_i \nabla f(\theta_i) - \zeta'_{n,k} \\ &= - (\gamma_k - \gamma_n) \nabla f(\theta_n) - \zeta_{n,k},\end{aligned}\tag{6.1}$$

$$f(\theta_k) - f(\theta_n) = -(\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 - \phi_{n,k}\tag{6.2}$$

for $0 \leq n \leq k$.

In this section, besides the quantities introduced in the previous paragraph, we rely on the following notation. For a compact set $Q \subset \mathbb{R}^{d_\theta}$, C_Q stands for an upper bound of $\|\nabla f(\cdot)\|$ on Q and for a Lipschitz constant of $\nabla f(\cdot)$ on the same set. \hat{A} is the set of accumulation points of $\{\theta_n\}_{n \geq 0}$, while

$$\hat{f} = \liminf_{n \rightarrow \infty} f(\theta_n).$$

\hat{B} and \hat{Q} are random sets defined by

$$\hat{B} = \bigcup_{\theta \in \hat{A}} \{\theta' \in \mathbb{R}^{d_\theta} : \|\theta' - \theta\| \leq \delta_\theta/2\}, \quad \hat{Q} = \text{cl}(\hat{B})$$

on event Λ , and by

$$\hat{B} = \hat{A}, \quad \hat{Q} = \hat{A}$$

outside Λ (δ_θ is specified in Remark 2.1). Overriding the definition of $\hat{\mu}$, \hat{p} , \hat{r} , in Theorem 2.2, we specify random quantities $\hat{\delta}$, $\hat{\mu}$, \hat{p} , \hat{r} , \hat{C} , \hat{M} as

$$\begin{aligned}\hat{\delta} &= \delta_{\hat{Q}, \hat{f}}, \quad \hat{\mu} = \mu_{\hat{Q}, \hat{f}}, \quad \hat{C} = C_{\hat{Q}}, \quad \hat{M} = \hat{M}_{\hat{Q}, \hat{f}}, \\ \hat{r} &= \begin{cases} 1/(2 - \hat{\mu}), & \text{if } \hat{\mu} < 2 \\ \infty, & \text{if } \hat{\mu} = 2 \end{cases}, \quad \hat{p} = \hat{\mu} \min\{r, \hat{r}\}\end{aligned}$$

on Λ ($\delta_{Q,a}$, $\mu_{Q,a}$, $M_{Q,a}$ are specified in Assumption 2.3), and as

$$\hat{\delta} = 1, \quad \hat{\mu} = 2, \quad \hat{C} = 1, \quad \hat{M} = 1, \quad \hat{r} = \infty, \quad \hat{p} = 2r$$

outside Λ (later, when Theorem 2.1 is proved, it will be clear that $\hat{\mu}, \hat{p}, \hat{r}$ specified here coincide with $\hat{\mu}, \hat{p}, \hat{r}$ defined in Theorem 2.2). Functions $u(\cdot), v(\cdot)$ are defined by

$$u(\theta) = f(\theta) - \hat{f}, \quad v(\theta) = \begin{cases} (f(\theta) - \hat{f})^{-1/\hat{p}}, & \text{if } f(\theta) > \hat{f} \\ 0, & \text{otherwise} \end{cases}$$

for $\theta \in \mathbb{R}^{d_\theta}$.

REMARK 6.1. *On event Λ , \hat{Q} is compact and satisfies $\hat{A} \subset \text{int}\hat{Q}$. Thus, $\hat{\delta}, \hat{p}, \hat{r}, \hat{C}, \hat{M}, v(\cdot)$ are well-defined on Λ (what happens with these quantities outside Λ does not affect the results presented in this section). On the other side, Assumption 2.3 implies*

$$|f(\theta) - \hat{f}| \leq \hat{M} \|\nabla f(\theta)\|^{\hat{\mu}} \quad (6.3)$$

on Λ for all $\theta \in \hat{Q}$ satisfying $|f(\theta) - \hat{f}| \leq \hat{\delta}$.

REMARK 6.2. *Regarding the notation, the following note is also in order: \sim symbol is used for a locally defined quantity, i.e., for a quantity whose definition holds only in the proof where such a quantity appears.*

LEMMA 6.1. *Let Assumptions 2.1 and 2.2 hold. Then, there exists an event $N_0 \in \mathcal{F}$ such that $P(N_0) = 0$ and*

$$\limsup_{n \rightarrow \infty} \gamma_n^r \max_{n \leq k \leq a(n,1)} \|\zeta'_{n,k}\| \leq \xi < \infty$$

on $\Lambda \setminus N_0$.

Proof. It is straightforward to verify

$$\zeta'_{n,k} = \sum_{i=n}^{k-1} (\gamma_i^{-r} - \gamma_{i+1}^{-r}) \left(\sum_{j=n}^i \alpha_j \gamma_j^r \xi_j \right) + \gamma_k^{-r} \sum_{i=n}^{k-1} \alpha_i \gamma_i^r \xi_i$$

for $0 \leq n < k$. Consequently,

$$\begin{aligned} \|\zeta'_{n,k}\| &\leq \left(\gamma_k^{-r} + \sum_{i=n}^{k-1} (\gamma_i^{-r} - \gamma_{i+1}^{-r}) \right) \max_{n \leq j < a(n,1)} \left\| \sum_{i=n}^j \alpha_i \gamma_i^r \xi_i \right\| \\ &= \gamma_n^{-r} \max_{n \leq j < a(n,1)} \left\| \sum_{i=n}^j \alpha_i \gamma_i^r \xi_i \right\| \end{aligned}$$

for $0 \leq n \leq k \leq a(n,1)$. Thus,

$$\gamma_n^r \|\zeta'_{n,k}\| \leq \max_{n \leq j < a(n,1)} \left\| \sum_{i=n}^j \alpha_i \gamma_i^r \xi_i \right\|$$

for $0 \leq n \leq k \leq a(n,1)$. Then, the lemma's assertion directly follows from Assumption 2.2. \square

LEMMA 6.2. *Suppose that Assumptions 2.1 – 2.3 hold. Then, there exist random quantities \hat{C}_1, \hat{t} (which are deterministic functions of \hat{C}) and for any real number $\varepsilon \in (0, \infty)$, there exists a non-negative integer-valued random quantity $\tau_{1,\varepsilon}$ such that*

the following is true: $1 \leq \hat{C}_1 < \infty$, $0 < \hat{t} < 1$, $0 \leq \tau_{1,\varepsilon} < \infty$ everywhere and

$$\max_{n \leq k \leq a(n, \hat{t})} \|\theta_k - \theta_n\| \leq \hat{C}_1 \left(\|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon) \right), \quad (6.4)$$

$$\max_{n \leq k \leq a(n, \hat{t})} (f(\theta_k) - f(\theta_n)) \leq \hat{C}_1 \left(\gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2 \right), \quad (6.5)$$

$$f(\theta_{a(n, \hat{t})}) - f(\theta_n) + \hat{t} \|\nabla f(\theta_n)\|^2 / 2 \leq \hat{C}_1 \left(\gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2 \right) \quad (6.6)$$

$$\begin{aligned} & 2 \left(f(\theta_{a(n, \hat{t})}) - f(\theta_n) \right) + \hat{t} \|\nabla f(\theta_n)\|^2 / 2 + \|\nabla f(\theta_n)\| \|\theta_{a(n, \hat{t})} - \theta_n\| \\ & \leq \hat{C}_1 \left(\gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2 \right) \end{aligned} \quad (6.7)$$

on $\Lambda \setminus N_0$ for $n > \tau_{1,\varepsilon}$.

Proof. Let $\tilde{C}_1 = 2\hat{C} \exp(\hat{C})$, $\tilde{C}_2 = 2\hat{C}\tilde{C}_1$, $\tilde{C}_3 = 2\hat{C}\tilde{C}_1^2 + \hat{C}_2$, $\tilde{C}_4 = \tilde{C}_2 + 2\tilde{C}_3$, while $\hat{C}_1 = \tilde{C}_4$, $\hat{t} = 1/(4\tilde{C}_4)$. Moreover, let $\varepsilon \in (0, \infty)$ be an arbitrary real number. Then, owing to Lemma 6.1 and the fact that $\gamma_{a(n, \hat{t})} - \gamma_n = \hat{t} + O(\alpha_{a(n, \hat{t})})$ for $n \rightarrow \infty$, it is possible to construct a non-negative integer-valued random quantity $\tau_{1,\varepsilon}$ such that $0 \leq \tau_{1,\varepsilon} < \infty$ everywhere and such that $\theta_n \in \hat{Q}$,

$$\gamma_{a(n, \hat{t})} - \gamma_n \geq 2\hat{t}/3, \quad (6.8)$$

$$\max_{n \leq k \leq a(n, 1)} \|\zeta'_{n,k}\| \leq \gamma_n^{-r}(\xi + \varepsilon) \quad (6.9)$$

on $\Lambda \setminus N_0$ for $n > \tau_{1,\varepsilon}$.

Let ω be an arbitrary sample from $\Lambda \setminus N_0$ (notice that all formulas which follow in the proof correspond to this sample). Since $\theta_n \in \hat{Q}$ for $n > \tau_{1,\varepsilon}$, (6.1), (6.9) yield

$$\begin{aligned} \|\nabla f(\theta_k)\| & \leq \|\nabla f(\theta_n)\| + \|\nabla f(\theta_k) - \nabla f(\theta_n)\| \\ & \leq \|\nabla f(\theta_n)\| + \hat{C} \|\theta_k - \theta_n\| \\ & \leq \|\nabla f(\theta_n)\| + \hat{C} \sum_{i=n}^{k-1} \alpha_i \|\nabla f(\theta_i)\| + \hat{C} \|\zeta'_{n,k}\| \\ & \leq \|\nabla f(\theta_n)\| + \hat{C} \gamma_n^{-r}(\xi + \varepsilon) + \hat{C} \sum_{i=n}^{k-1} \alpha_i \|\nabla f(\theta_i)\| \end{aligned}$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, 1)$. Then, Bellman-Gronwall inequality implies

$$\begin{aligned} \|\nabla f(\theta_k)\| & \leq \left(\|\nabla f(\theta_n)\| + \hat{C} \gamma_n^{-r}(\xi + \varepsilon) \right) \exp \left(\hat{C}(\gamma_k - \gamma_n) \right) \\ & \leq \hat{C} \exp(\hat{C}) \left(\|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon) \right) \end{aligned}$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, 1)$ (notice that $\gamma_k - \gamma_n \leq \gamma_{a(n, 1)} - \gamma_n \leq 1$ when $n \leq k \leq a(n, 1)$). Consequently, (6.9) gives

$$\begin{aligned} \|\theta_k - \theta_n\| & \leq \sum_{i=n}^{k-1} \alpha_i \|\nabla f(\theta_i)\| + \|\zeta'_{n,k}\| \\ & \leq \hat{C} \exp(\hat{C}) \left(\|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon) \right) (\gamma_k - \gamma_n) + \gamma_n^{-r}(\xi + \varepsilon) \\ & \leq \tilde{C}_1 \left((\gamma_k - \gamma_n) \|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon) \right) \end{aligned} \quad (6.10)$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, 1)$. Therefore, (6.9) yields

$$\begin{aligned} \|\zeta_{n,k}\| &\leq \|\zeta'_{n,k}\| + \hat{C} \sum_{i=n}^{k-1} \alpha_i \|\theta_i - \theta_n\| \\ &\leq \gamma_n^{-r}(\xi + \varepsilon) + \hat{C} \tilde{C}_1 ((\gamma_k - \gamma_n) \|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon)) (\gamma_k - \gamma_n) \\ &\leq \tilde{C}_2 ((\gamma_k - \gamma_n)^2 \|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon)) \end{aligned} \quad (6.11)$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, 1)$. Thus,

$$\begin{aligned} |\phi_{n,k}| &\leq \|\nabla f(\theta_n)\| \|\zeta_{n,k}\| + \hat{C} \|\theta_k - \theta_n\|^2 \\ &\leq \tilde{C}_2 ((\gamma_k - \gamma_n)^2 \|\nabla f(\theta_n)\|^2 + \gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon)) \\ &\quad + \hat{C} \tilde{C}_1^2 ((\gamma_k - \gamma_n) \|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon))^2 \\ &\leq \tilde{C}_3 ((\gamma_k - \gamma_n)^2 \|\nabla f(\theta_n)\|^2 + \gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2) \end{aligned} \quad (6.12)$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, 1)$.

Owing to (6.2), (6.12), we have

$$\begin{aligned} f(\theta_k) - f(\theta_n) &\leq -(\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 + |\phi_{n,k}| \\ &\leq -\left(1 - \tilde{C}_3(\gamma_k - \gamma_n)\right) (\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 \\ &\quad + \tilde{C}_3 (\gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2) \end{aligned} \quad (6.13)$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, 1)$. Since

$$\tilde{C}_3(\gamma_k - \gamma_n) \leq \tilde{C}_4(\gamma_k - \gamma_n) \leq \tilde{C}_4(\gamma_{a(n,\hat{t})} - \gamma_n) \leq \tilde{C}_4 \hat{t} \leq 1/4 \quad (6.14)$$

for $0 \leq n \leq k \leq a(n, \hat{t})$, (6.13) yields

$$\begin{aligned} f(\theta_k) - f(\theta_n) &\leq -3(\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 / 4 \\ &\quad + \tilde{C}_3 (\gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2) \end{aligned} \quad (6.15)$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, \hat{t})$. As an immediate consequence of (6.8), (6.10), (6.15) we get that (6.4) - (6.6) hold for $n > \tau_{1,\varepsilon}$ (notice that $\gamma_k - \gamma_n \leq 1$ for $n \leq k \leq a(n, 1)$).

Due to (6.1), we have

$$\begin{aligned} (\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 &= \|\nabla f(\theta_n)\| \|(\gamma_k - \gamma_n) \nabla f(\theta_n)\| \\ &= \|\nabla f(\theta_n)\| \|\theta_k - \theta_n + \zeta_{n,k}\| \end{aligned}$$

for $0 \leq n \leq k$. Combining this with (6.2), (6.12) and the first part of (6.11), we get

$$\begin{aligned} 2(f(\theta_k) - f(\theta_n)) &= -\|\nabla f(\theta_n)\| \|\theta_k - \theta_n + \zeta_{n,k}\| - (\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 - 2\phi_{n,k} \\ &\leq -\|\nabla f(\theta_n)\| \|\theta_k - \theta_n\| - (\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 \\ &\quad + \|\nabla f(\theta_n)\| \|\zeta_{n,k}\| + 2|\phi_{n,k}| \\ &\leq -\|\nabla f(\theta_n)\| \|\theta_k - \theta_n\| - (\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 \\ &\quad + \tilde{C}_4(\gamma_k - \gamma_n)^2 \|\nabla f(\theta_n)\|^2 \\ &\quad + \tilde{C}_4 (\gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2) \\ &= -\|\nabla f(\theta_n)\| \|\theta_k - \theta_n\| - \left(1 - \tilde{C}_4(\gamma_k - \gamma_n)\right) (\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2 \\ &\quad + \tilde{C}_4 (\gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2) \end{aligned}$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, 1)$. Consequently, (6.14) yields

$$\begin{aligned} 2(f(\theta_k) - f(\theta_n)) &\leq -\|\nabla f(\theta_n)\| \|\theta_k - \theta_n\| - 3(\gamma_k - \gamma_n) \|\nabla f(\theta_n)\|^2/4 \\ &\quad + \tilde{C}_4 (\gamma_n^{-r} \|\nabla f(\theta_n)\| (\xi + \varepsilon) + \gamma_n^{-2r} (\xi + \varepsilon)^2) \end{aligned}$$

for $\tau_{1,\varepsilon} < n \leq k \leq a(n, \hat{t})$. Then, (6.8) implies that (6.7) is true for $n > \tau_{1,\varepsilon}$. \square

LEMMA 6.3. *Suppose that Assumptions 2.1 – 2.3 hold. Then, $\lim_{n \rightarrow \infty} \nabla f(\theta_n) = 0$ on $\Lambda \setminus N_0$.*

Proof. The lemma's assertion is proved by contradiction. We assume that $\limsup_{n \rightarrow \infty} \|\nabla f(\theta_n)\| > 0$ for some sample $\omega \in \Lambda \setminus N_0$ (notice that all formulas which follow in the proof correspond to this sample). Then, there exists $a \in (0, \infty)$ and an increasing sequence $\{l_k\}_{k \geq 0}$ (both depending on ω) such that $\liminf_{k \rightarrow \infty} \|\nabla f(\theta_{l_k})\| > a$. Since $\liminf_{k \rightarrow \infty} f(\theta_{a(l_k, \hat{t})}) \geq \hat{f}$, Lemma 6.2 (inequality (6.6)) gives

$$\begin{aligned} \hat{f} - \liminf_{k \rightarrow \infty} f(\theta_{l_k}) &\leq \limsup_{k \rightarrow \infty} (f(\theta_{a(l_k, \hat{t})}) - f(\theta_{l_k})) \\ &\leq -(\hat{t}/2) \liminf_{k \rightarrow \infty} \|\nabla f(\theta_{l_k})\|^2 \\ &\leq -a^2 \hat{t}/2. \end{aligned}$$

Therefore, $\liminf_{k \rightarrow \infty} f(\theta_{l_k}) \geq \hat{f} + a\hat{t}^2/2$. Consequently, there exist $b, c \in \mathbb{R}$ (depending on ω) such that $\hat{f} < b < c < \hat{f} + a\hat{t}^2/2$, $b < \hat{f} + \hat{\delta}$ and $\limsup_{n \rightarrow \infty} f(\theta_n) > c$. Thus, there exist sequences $\{m_k\}_{k \geq 0}$, $\{n_k\}_{k \geq 0}$ (depending on ω) with the following properties: $m_k < n_k < m_{k+1}$, $f(\theta_{m_k}) < b$, $f(\theta_{n_k}) > c$ and

$$\max_{m_k < n \leq n_k} f(\theta_n) \geq b \quad (6.16)$$

for $k \geq 0$. Then, Lemma 6.2 (inequality (6.5)) implies

$$\limsup_{k \rightarrow \infty} (f(\theta_{m_k+1}) - f(\theta_{m_k})) \leq 0, \quad (6.17)$$

$$\limsup_{k \rightarrow \infty} \max_{m_k \leq n \leq a(m_k, \hat{t})} (f(\theta_n) - f(\theta_{m_k})) \leq 0. \quad (6.18)$$

Since

$$b > f(\theta_{m_k}) = f(\theta_{m_k+1}) - (f(\theta_{m_k+1}) - f(\theta_{m_k})) \geq b - (f(\theta_{m_k+1}) - f(\theta_{m_k}))$$

for $k \geq 0$, (6.17) yields $\lim_{k \rightarrow \infty} f(\theta_{m_k}) = b$. As $f(\theta_{n_k}) - f(\theta_{m_k}) > c - b$ for $k \geq 0$, (6.18) implies $a(m_k, \hat{t}) < n_k$ for all, but infinitely many k (otherwise, $\liminf_{k \rightarrow \infty} (f(\theta_{n_k}) - f(\theta_{m_k})) \leq 0$ would follow from (6.18)). Consequently, $\liminf_{k \rightarrow \infty} f(\theta_{a(m_k, \hat{t})}) \geq b$ (due to (6.16)), while Lemma 6.2 (inequality (6.6)) gives

$$\begin{aligned} 0 &\leq \limsup_{k \rightarrow \infty} f(\theta_{a(m_k, \hat{t})}) - b = \limsup_{k \rightarrow \infty} (f(\theta_{a(m_k, \hat{t})}) - f(\theta_{m_k})) \\ &\leq -(\hat{t}/2) \liminf_{k \rightarrow \infty} \|\nabla f(\theta_{m_k})\|^2. \end{aligned}$$

Therefore, $\lim_{k \rightarrow \infty} \|\nabla f(\theta_{m_k})\| = 0$. Moreover, there exists $k_0 \geq 0$ (depending on ω) such that $\theta_{m_k} \in \hat{Q}$ and $f(\theta_{m_k}) \geq (\hat{f} + b)/2$ for $k \geq k_0$ (notice that $\lim_{k \rightarrow \infty} f(\theta_{m_k}) = b > (\hat{f} + b)/2$). Consequently, $\theta_{m_k} \in \hat{Q}$ and $0 < (b - \hat{f})/2 \leq f(\theta_{m_k}) - \hat{f} \leq \hat{\delta}$ for $k \geq k_0$

(notice that $f(\theta_{m_k}) < b < \hat{f} + \hat{\delta}$ for $k \geq 0$). Then, owing to (6.3) (i.e., to Assumption 3.3), we have

$$0 < (b - \hat{f})/2 \leq f(\theta_{m_k}) - \hat{f} \leq \hat{M} \|\nabla f(\theta_{m_k})\|^{\hat{\mu}}$$

for $k \geq k_0$. However, this directly contradicts the fact $\lim_{k \rightarrow \infty} \|\nabla f(\theta_{m_k})\| = 0$. Hence, $\lim_{n \rightarrow \infty} \nabla f(\theta_n) = 0$ on $\Lambda \setminus N_0$. \square

LEMMA 6.4. *Suppose that Assumptions 2.1 – 2.3 hold. Then, $\lim_{n \rightarrow \infty} f(\theta_n) = \hat{f}$ on $\Lambda \setminus N_0$.*

Proof. We use contradiction to prove the lemma's assertion: Suppose that $\hat{f} < \limsup_{n \rightarrow \infty} f(\theta_n)$ for some sample $\omega \in \Lambda \setminus N_0$ (notice that all formulas which follow in the proof correspond to this sample). Then, there exists $a \in \mathbb{R}$ (depending on ω) such that $\hat{f} < a < \hat{f} + \hat{\delta}$ and $\limsup_{n \rightarrow \infty} f(\theta_n) > a$. Thus, there exists an increasing sequence $\{n_k\}_{k \geq 0}$ (depending on ω) such that $f(\theta_{n_k}) < a$ and $f(\theta_{n_{k+1}}) \geq a$ for $k \geq 0$. On the other side, Lemma 6.2 (inequality (6.5)) implies

$$\limsup_{k \rightarrow \infty} (f(\theta_{n_{k+1}}) - f(\theta_{n_k})) \leq 0. \quad (6.19)$$

Since

$$a > f(\theta_{n_k}) = f(\theta_{n_{k+1}}) - (f(\theta_{n_{k+1}}) - f(\theta_{n_k})) \geq a - (f(\theta_{n_{k+1}}) - f(\theta_{n_k}))$$

for $k \geq 0$, (6.19) yields $\lim_{k \rightarrow \infty} f(\theta_{n_k}) = a$. Moreover, there exists $k_0 \geq 0$ (depending on ω) such that $\theta_{n_k} \in \hat{Q}$ and $f(\theta_{n_k}) \geq (\hat{f} + a)/2$ for $k \geq k_0$ (notice that $\lim_{k \rightarrow \infty} f(\theta_{n_k}) = a > (\hat{f} + a)/2$). Thus, $\theta_{n_k} \in \hat{Q}$ and $0 < (a - \hat{f})/2 \leq f(\theta_{n_k}) - \hat{f} \leq \hat{\delta}$ for $k \geq k_0$ (notice that $f(\theta_{n_k}) < a < \hat{f} + \hat{\delta}$ for $k \geq 0$). Then, due to (6.3) (i.e., to Assumption 2.3), we have

$$0 < (a - \hat{f})/2 \leq f(\theta_{n_k}) - \hat{f} \leq \hat{M} \|\nabla f(\theta_{n_k})\|^{\hat{\mu}}$$

for $k \geq k_0$. However, this directly contradicts the fact $\lim_{n \rightarrow \infty} \nabla f(\theta_n) = 0$. Hence, $\lim_{n \rightarrow \infty} f(\theta_n) = \hat{f}$ on $\Lambda \setminus N_0$. \square

LEMMA 6.5. *Suppose that Assumptions 2.1 – 2.3 hold. Then, there exist random quantities \hat{C}_2, \hat{C}_3 (which are deterministic functions of $\hat{p}, \hat{C}, \hat{M}$) and for any real number $\varepsilon \in (0, \infty)$, there exists a non-negative integer-valued random quantity $\tau_{2,\varepsilon}$ such that the following is true: $1 \leq \hat{C}_2, \hat{C}_3 < \infty$, $0 \leq \tau_{2,\varepsilon} < \infty$ everywhere and*

$$\left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) + \hat{t} \|\nabla f(\theta_n)\|^2/4 \right) I_{A_{n,\varepsilon}} \leq 0, \quad (6.20)$$

$$\left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) + (\hat{t}/\hat{C}_3) u(\theta_n) \right) I_{B_{n,\varepsilon}} \leq 0, \quad (6.21)$$

$$\left(v(\theta_{a(n,\hat{t})}) - v(\theta_n) - (\hat{t}/\hat{C}_3)(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} \right) I_{C_{n,\varepsilon}} \geq 0 \quad (6.22)$$

on $\Lambda \setminus N_0$ for $n \geq \tau_{2,\varepsilon}$, where

$$A_{n,\varepsilon} = \left\{ \gamma_n^{\hat{p}} |u(\theta_n)| \geq \hat{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}} \right\} \cup \left\{ \gamma_n^{\hat{p}} \|\nabla f(\theta_n)\|^2 \geq \hat{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}} \right\},$$

$$B_{n,\varepsilon} = \left\{ \gamma_n^{\hat{p}} u(\theta_n) \geq \hat{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}} \right\} \cap \{\hat{\mu} = 2\},$$

$$C_{n,\varepsilon} = \left\{ \gamma_n^{\hat{p}} u(\theta_n) \geq \hat{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}} \right\} \cap \left\{ u(\theta_{a(n,\hat{t})}) > 0 \right\} \cap \{\hat{\mu} < 2\}.$$

REMARK 6.3. *Inequalities (6.20) – (6.22) can be represented in the following equivalent form: Relations*

$$\begin{aligned} & \left(\gamma_n^{\hat{p}} |u(\theta_n)| \geq \hat{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}} \vee \gamma_n^{\hat{p}} \|\nabla f(\theta_n)\|^2 \geq \hat{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}} \right) \wedge n > \tau_{2,\varepsilon} \\ \implies & u(\theta_{a(n,\hat{t})}) \leq u(\theta_n) - \hat{t} \|\nabla f(\theta_n)\|^2 / 4, \end{aligned} \quad (6.23)$$

$$\begin{aligned} & \gamma_n^{\hat{p}} u(\theta_n) \geq \hat{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}} \wedge \hat{\mu} = 2 \wedge n > \tau_{2,\varepsilon} \\ \implies & u(\theta_{a(n,\hat{t})}) \leq \left(1 - \hat{t}/\hat{C}_3\right) u(\theta_n), \end{aligned} \quad (6.24)$$

$$\begin{aligned} & \gamma_n^{\hat{p}} u(\theta_n) \geq \hat{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}} \wedge u(\theta_{a(n,\hat{t})}) > 0 \wedge \hat{\mu} < 2 \wedge n > \tau_{2,\varepsilon} \\ \implies & v(\theta_{a(n,\hat{t})}) \geq v(\theta_n) + (\hat{t}/\hat{C}_3)(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} \end{aligned} \quad (6.25)$$

are true on $\Lambda \setminus N_0$.

Proof. Let $\tilde{C} = 8\hat{C}_1/\hat{t}$, $\hat{C}_2 = \tilde{C}^2\hat{M}$ and $\hat{C}_3 = 4\hat{p}\hat{M}^2$. Moreover, let $\varepsilon \in (0, \infty)$ be an arbitrary real number. Then, owing to Lemma 6.1 and 6.4, it is possible to construct a non-negative inter-valued random quantity $\tau_{2,\varepsilon}$ such that $\tau_{1,\varepsilon} \leq \tau_{2,\varepsilon} < \infty$ everywhere and such that $\theta_n \in \hat{Q}$, $|u(\theta_n)| \leq \hat{\delta}$,

$$\gamma_n^{-\hat{p}/2}(\varphi_\varepsilon(\xi))^{\hat{\mu}/2} \geq \gamma_n^{-r}(\xi + \varepsilon), \quad (6.26)$$

$$\gamma_n^{-\hat{p}/\hat{\mu}}\varphi_\varepsilon(\xi) \geq \gamma_n^{-r}(\xi + \varepsilon) \quad (6.27)$$

on $\Lambda \setminus N_0$ for $n > \tau_{2,\varepsilon}$.⁴ Since $\tau_{2,\varepsilon} \geq \tau_{1,\varepsilon}$ on $\Lambda \setminus N_0$, Lemma 6.2 (inequality (6.6)) yields

$$u(\theta_{a(n,\hat{t})}) - u(\theta_n) \leq -\hat{t} \|\nabla f(\theta_n)\|^2 / 2 + \hat{C}_1 \left(\gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2 \right) \quad (6.28)$$

on $\Lambda \setminus N_0$ for $n > \tau_{2,\varepsilon}$. As $\theta_n \in \hat{Q}$ and $|u(\theta_n)| \leq \hat{\delta}$ on $\Lambda \setminus N_0$ for $n > \tau_{2,\varepsilon}$, (6.3) (i.e., Assumption 2.3) implies

$$|u(\theta_n)| \leq \hat{M} \|\nabla f(\theta_n)\|^{\hat{\mu}} \quad (6.29)$$

on $\Lambda \setminus N_0$ for $n > \tau_{2,\varepsilon}$.

Let ω be an arbitrary sample from $\Lambda \setminus N_0$ (notice that all formulas which follow in the proof correspond to this sample). First, we show (6.20). We proceed by contradiction: Suppose that (6.20) is violated for some $n > \tau_{2,\varepsilon}$. Therefore,

$$u(\theta_{a(n,\hat{t})}) - u(\theta_n) > -\hat{t} \|\nabla f(\theta_n)\|^2 / 4 \quad (6.30)$$

and at least one of the following two inequalities is true:

$$|u(\theta_n)| \geq \hat{C}_2 \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}}, \quad (6.31)$$

$$\|\nabla f(\theta_n)\|^2 \geq \hat{C}_2 \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}}. \quad (6.32)$$

⁴ To conclude that (6.26) holds on $\Lambda \setminus N_0$ for all but finitely many n , notice that $\hat{p}/2 < \min\{r, \hat{r}\} \leq r$ when $\hat{\mu} < 2$ and that the left and right hand sides of the inequality in (6.26) are equal when $\hat{\mu} = 2$. In order to deduce that (6.27) is true on $\Lambda \setminus N_0$ for all but finitely many n , notice that $\hat{p}/\hat{\mu} = r$, $\varphi_\varepsilon(\xi) \geq \xi + \varepsilon$ when $r \leq \hat{r}$ and that $\hat{p}/\hat{\mu} = \hat{r} < r$ when $r > \hat{r}$.

If (6.31) holds, then (6.27), (6.29) imply

$$\|\nabla f(\theta_n)\| \geq (|u(\theta_n)|/\hat{M})^{1/\hat{\mu}} \geq (\hat{C}_2/\hat{M})^{1/\hat{\mu}} \gamma_n^{-\hat{p}/\hat{\mu}} \varphi_\varepsilon(\xi) \geq \tilde{C} \gamma_n^{-r}(\xi + \varepsilon)$$

(notice that $(\hat{C}_2/\hat{M})^{1/\hat{\mu}} = \tilde{C}^{2/\hat{\mu}} \geq \tilde{C}$ owing to $\hat{\mu} \leq 2$). On the other side, if (6.32) is satisfied, then (6.26) yields

$$\|\nabla f(\theta_n)\| \geq \hat{C}_2^{1/2} \gamma_n^{-\hat{p}/2} (\varphi_\varepsilon(\xi))^{\hat{\mu}/2} \geq \tilde{C} \gamma_n^{-r}(\xi + \varepsilon).$$

Thus, as a result of one of (6.31), (6.32), we get

$$\|\nabla f(\theta_n)\| \geq \tilde{C} \gamma_n^{-r}(\xi + \varepsilon).$$

Consequently,

$$\begin{aligned} \hat{t} \|\nabla f(\theta_n)\|^2/8 &\geq (\tilde{C}\hat{t}/8) \gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon) = \hat{C}_1 \gamma_n^{-r} \|\nabla f(\theta_n)\|(\xi + \varepsilon), \\ \hat{t} \|\nabla f(\theta_n)\|^2/8 &\geq (\tilde{C}^2\hat{t}/8) \gamma_n^{-2r}(\xi + \varepsilon)^2 \geq \hat{C}_1 \gamma_n^{-2r}(\xi + \varepsilon)^2 \end{aligned}$$

(notice that $\tilde{C}\hat{t}/8 = \hat{C}_1$, $\tilde{C}^2\hat{t}/8 \geq \tilde{C}\hat{t}/8 = \hat{C}_1$). Combining this with (6.28), we get

$$u(\theta_{a(n,\hat{t})}) - u(\theta_n) \leq -\hat{t} \|\nabla f(\theta_n)\|^2/4, \quad (6.33)$$

which directly contradicts (6.30). Hence, (6.20) is true for $n > \tau_{2,\varepsilon}$. Then, as a result of (6.29) and the fact that $B_{n,\varepsilon} \subseteq A_{n,\varepsilon}$ for $n \geq 0$, we get

$$\begin{aligned} &\left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) + (\hat{t}/\hat{C}_3) u(\theta_n) \right) I_{B_{n,\varepsilon}} \\ &\leq \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) + (\hat{M}\hat{t}/\hat{C}_3) \|\nabla f(\theta_n)\|^2 \right) I_{B_{n,\varepsilon}} \\ &\leq \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) + \hat{t} \|\nabla f(\theta_n)\|^2/4 \right) I_{B_{n,\varepsilon}} \leq 0 \end{aligned}$$

for $n > \tau_{2,\varepsilon}$ (notice that $u(\theta_n) > 0$ on $B_{n,\varepsilon}$ for each $n \geq 0$; also notice that $\hat{C}_3 \geq 4\hat{M}$). Thus, (6.21) is true for $n > \tau_{2,\varepsilon}$.

Now, let us prove (6.22). To do so, we again use contradiction: Suppose that (6.21) does not hold for some $n > \tau_{2,\varepsilon}$. Consequently, we have $\hat{\mu} < 2$, $u(\theta_{a(n,\hat{t})}) > 0$ and

$$\gamma_n^{\hat{p}} u(\theta_n) \geq \hat{C}_2 (\varphi_\varepsilon(\xi))^{\hat{\mu}} > 0, \quad (6.34)$$

$$v(\theta_{a(n,\hat{t})}) - v(\theta_n) < (\hat{t}/\hat{C}_3) (\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}}. \quad (6.35)$$

Combining (6.34) with (already proved) (6.20), we get (6.33), while $\hat{\mu} < 2$ implies

$$2/\hat{\mu} = 1 + 1/(\hat{\mu}\hat{r}) \leq 1 + 1/\hat{p} \quad (6.36)$$

(notice that $\hat{r} = 1/(2 - \hat{\mu})$ owing to $\hat{\mu} < 2$; also notice that $\hat{p} = \hat{\mu} \min\{r, \hat{r}\} \leq \hat{\mu}\hat{r}$). As $0 < u(\theta_n) \leq \hat{\delta} \leq 1$ (due to (6.34) and the definition of $\tau_{2,\varepsilon}$), inequalities (6.29), (6.36) yield

$$\|\nabla f(\theta_n)\|^2 \geq \left(u(\theta_n)/\hat{M} \right)^{2/\hat{\mu}} \geq (u(\theta_n))^{1+1/\hat{p}} / \hat{M}^2 \quad (6.37)$$

(notice that $\hat{M}^{2/\hat{\mu}} \leq \hat{M}^2$ due to $\hat{\mu} < 2$, $\hat{M} \geq 1$). Since $\|\nabla f(\theta_n)\| > 0$ and $0 < u(\theta_{a(n,\hat{t})}) < u(\theta_n)$ (due to (6.29), (6.33)), inequalities (6.33), (6.37) give

$$\begin{aligned} \frac{\hat{t}}{4} &\leq \frac{u(\theta_n) - u(\theta_{a(n,\hat{t})})}{\|\nabla f(\theta_n)\|^2} \leq \hat{M}^2 \frac{u(\theta_n) - u(\theta_{a(n,\hat{t})})}{(u(\theta_n))^{1+1/\hat{p}}} \\ &= \hat{M}^2 \int_{u(\theta_{a(n,\hat{t})})}^{u(\theta_n)} \frac{du}{(u(\theta_n))^{1+1/\hat{p}}} \\ &\leq \hat{M}^2 \int_{u(\theta_{a(n,\hat{t})})}^{u(\theta_n)} \frac{du}{u^{1+1/\hat{p}}} \\ &= \hat{p}\hat{M}^2 \left(v(\theta_{a(n,\hat{t})}) - v(\theta_n) \right). \end{aligned}$$

Therefore,

$$v(\theta_{a(n,\hat{t})}) - v(\theta_n) \geq \hat{t}/(4\hat{p}\hat{M}^2) = (\hat{t}/\hat{C}_3),$$

which directly contradicts (6.35). Thus, (6.22) is satisfied for $n > \tau_{2,\varepsilon}$. \square

LEMMA 6.6. *Suppose that Assumptions 2.1 – 2.3 hold. Then, there exists a random quantity \hat{C}_4 (which is a deterministic function of \hat{C}) such that the following is true: $1 \leq \hat{C}_4 < \infty$ everywhere and*

$$\|\theta_{a(n,\hat{t})} - \theta_n\| \leq -\gamma_n^s \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right) (\phi_{\varepsilon,s}(\xi))^{-1} + \hat{C}_4 \gamma_n^{-s} \phi_{\varepsilon,s}(\xi) \quad (6.38)$$

on $\Lambda \setminus N_0$ for $n > \tau_{1,\varepsilon}$ and any $\varepsilon \in (0, \infty)$, $s \in (1, r]$, where

$$\phi_{\varepsilon,s}(\xi) = \begin{cases} 1 + \xi + \varepsilon, & \text{if } r = s, r > \hat{r} \\ \varphi_\varepsilon(\xi), & \text{otherwise} \end{cases}.$$

Proof. Let $\varepsilon \in (0, \infty)$, $s \in (1, r]$ be arbitrary real numbers, while $\hat{C}_4 = 10\hat{C}_1^2/\hat{t}$. Moreover, let ω be an arbitrary sample from $\Lambda \setminus N_0$ (notice that all formulas which follow in the proof correspond to this sample), while $n > \tau_{1,\varepsilon}$ is an arbitrary integer. To prove (6.38), we consider separately the cases $\|\nabla f(\theta_n)\| \geq (4\hat{C}_1/\hat{t})\gamma_n^{-s}\phi_{\varepsilon,s}(\xi)$ and $\|\nabla f(\theta_n)\| \leq (4\hat{C}_1/\hat{t})\gamma_n^{-s}\phi_{\varepsilon,s}(\xi)$.

Case $\|\nabla f(\theta_n)\| \geq (4\hat{C}_1/\hat{t})\gamma_n^{-s}\phi_{\varepsilon,s}(\xi)$: Since $s \leq r$ and $\phi_{\varepsilon,s}(\xi) \geq \xi + \varepsilon$, we have

$$\|\nabla f(\theta_n)\| \geq (4\hat{C}_1/\hat{t})\gamma_n^{-r}(\xi + \varepsilon).$$

Therefore,

$$\begin{aligned} (\hat{t}/4)\|\nabla f(\theta_n)\|^2 &\geq \hat{C}_1\gamma_n^{-r}\|\nabla f(\theta_n)\|(\xi + \varepsilon), \\ (\hat{t}/4)\|\nabla f(\theta_n)\|^2 &\geq (4\hat{C}_1^2/\hat{t})\gamma_n^{-2r}(\xi + \varepsilon)^2 \geq \hat{C}_1\gamma_n^{-2r}(\xi + \varepsilon)^2. \end{aligned}$$

Then, Lemma 6.2 (inequality (6.7)) yields

$$\begin{aligned} \|\nabla f(\theta_n)\| \|\theta_{a(n,\hat{t})} - \theta_n\| &\leq -2 \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right) - \hat{t}\|\nabla f(\theta_n)\|^2/2 \\ &\quad + \hat{C}_1 \left(\gamma_n^{-r}\|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2 \right) \\ &\leq -2 \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right). \end{aligned}$$

Consequently,

$$\begin{aligned}
\|\theta_{a(n,\hat{t})} - \theta_n\| &\leq -2\|\nabla f(\theta_n)\|^{-1} \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right) \\
&\leq -(2\hat{C}_1/\hat{t})^{-1} \gamma_n^s \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right) (\phi_{\varepsilon,s}(\xi))^{-1} \\
&\leq -\gamma_n^s \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right) (\phi_{\varepsilon,s}(\xi))^{-1} + \hat{C}_4 \gamma_n^{-s} \phi_{\varepsilon,s}(\xi).
\end{aligned}$$

Hence, (6.38) is true when $\|\nabla f(\theta_n)\| \geq (4\hat{C}_1/\hat{t})\gamma_n^{-s}\phi_{\varepsilon,s}(\xi)$.

Case $\|\nabla f(\theta_n)\| \leq (4\hat{C}_1/\hat{t})\gamma_n^{-s}\phi_{\varepsilon,s}(\xi)$: As $s \leq r$ and $\phi_{\varepsilon,s}(\xi) \geq \xi + \varepsilon$, Lemma 6.2 (inequalities (6.4), (6.5)) implies

$$\begin{aligned}
\|\theta_{a(n,\hat{r})} - \theta_n\| &\leq \hat{C}_1 (\|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon)) \\
&\leq (\hat{C}_4/2)\gamma_n^{-s}\phi_{\varepsilon,s}(\xi), \\
u(\theta_{a(n,\hat{t})}) - u(\theta_n) &\leq \hat{C}_1 (\gamma_n^{-r}\|\nabla f(\theta_n)\|(\xi + \varepsilon) + \gamma_n^{-2r}(\xi + \varepsilon)^2) \\
&\leq (\hat{C}_4/2)\gamma_n^{-2s}(\phi_{\varepsilon,s}(\xi))^2.
\end{aligned}$$

Combining this, we get

$$\begin{aligned}
\|\theta_{a(n,\hat{t})} - \theta_n\| &\leq -\gamma_n^s \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right) (\phi_{\varepsilon,s}(\xi))^{-1} \\
&\quad + \gamma_n^s \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right) (\phi_{\varepsilon,s}(\xi))^{-1} + (\hat{C}_4/2)\gamma_n^{-s}\phi_{\varepsilon,s}(\xi) \\
&\leq -\gamma_n^s \left(u(\theta_{a(n,\hat{t})}) - u(\theta_n) \right) (\phi_{\varepsilon,s}(\xi))^{-1} + \hat{C}_4 \gamma_n^{-s}\phi_{\varepsilon,s}(\xi).
\end{aligned}$$

Thus, (6.38) holds when $\|\nabla f(\theta_n)\| \leq (4\hat{C}_1/\hat{t})\gamma_n^{-s}\phi_{\varepsilon,s}(\xi)$. \square

LEMMA 6.7. *Suppose that Assumptions 2.1 – 2.3 hold. Then,*

$$u(\theta_n) \geq -\hat{C}_2 \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \quad (6.39)$$

on $\Lambda \setminus N_0$ for $n > \tau_{2,\varepsilon}$ and any $\varepsilon \in (0, \infty)$. Furthermore, there exists a random quantity $\hat{C}_5 \in [1, \infty)$ (which is a deterministic function of \hat{p} , \hat{C} , \hat{M}) such that the following is true: $1 \leq \hat{C}_5 < \infty$ everywhere and

$$\|\nabla f(\theta_n)\|^2 \leq \hat{C}_5 (\psi(u(\theta_n)) + \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}}) \quad (6.40)$$

on $\Lambda \setminus N_0$ for $n > \tau_{2,\varepsilon}$ and any $\varepsilon \in (0, \infty)$, where function $\psi(\cdot)$ is defined by $\psi(x) = x \mathbf{I}_{(0,\infty)}(x)$, $x \in \mathbb{R}$.

Proof. Let $\hat{C}_5 = 4\hat{C}_2/\hat{t}$, while $\varepsilon \in (0, \infty)$ is an arbitrary real number. Moreover, ω is an arbitrary sample from $\Lambda \setminus N_0$ (notice that all formulas which follow in the proof correspond to this sample).

First, we prove (6.39). To do so, we use contradiction: Assume that (6.39) is not satisfied for some $n > \tau_{2,\varepsilon}$. Define $\{n_k\}_{k \geq 0}$ recursively by $n_0 = n$ and $n_k = a(n_{k-1}, \hat{t})$ for $k \geq 1$. Let us show by induction that $\{u(\theta_{n_k})\}_{k \geq 0}$ is non-increasing: Suppose that $u(\theta_{n_l}) \leq u(\theta_{n_{l-1}})$ for $0 \leq l \leq k$. Consequently,

$$u(\theta_{n_k}) \leq u(\theta_{n_0}) \leq -\hat{C}_2 \gamma_{n_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \leq -\hat{C}_2 \gamma_{n_k}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}}$$

(notice that $\{\gamma_n\}_{n \geq 0}$ is increasing). Then, Lemma 6.5 (relations (6.20), (6.23)) yields

$$u(\theta_{n_{k+1}}) - u(\theta_{n_k}) \leq -\hat{t}\|\nabla f(\theta_{n_k})\|^2/4 \leq 0,$$

i.e., $u(\theta_{n_{k+1}}) \leq u(\theta_{n_k})$. Thus, $\{u(\theta_{n_k})\}_{k \geq 0}$ is non-increasing. Therefore,

$$\limsup_{n \rightarrow \infty} u(\theta_{n_k}) \leq u(\theta_{n_0}) < 0.$$

However, this is not possible, as $\lim_{n \rightarrow \infty} u(\theta_n) = 0$ (due to Lemma 6.4). Hence, (6.39) indeed holds for $n > \tau_{2,\varepsilon}$.

Now, (6.40) is demonstrated. Again, we proceed by contradiction: Suppose that (6.40) is violated for some $n > \tau_{2,\varepsilon}$. Consequently,

$$\|\nabla f(\theta_n)\|^2 \geq \hat{C}_5 \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \geq \hat{C}_2 \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}}$$

(notice that $\hat{C}_5 \geq \hat{C}_2$), which, together with Lemma 6.5 (relations (6.20), (6.23)), yields

$$u(\theta_{a(n,\hat{t})}) - u(\theta_n) \leq -\hat{t} \|\nabla f(\theta_n)\|^2 / 4.$$

Then, (6.39) implies

$$\begin{aligned} \|\nabla f(\theta_n)\|^2 &\leq (4/\hat{t}) \left(u(\theta_n) - u(\theta_{a(n,\hat{t})}) \right) \\ &\leq (4/\hat{t}) \left(\psi(u(\theta_n)) + \hat{C}_2 \gamma_{a(n,\hat{t})}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \right) \\ &\leq \hat{C}_5 \left(\psi(u(\theta_n)) + \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \right). \end{aligned}$$

However, this directly contradicts our assumption that n violates (6.40). Thus, (6.40) is indeed satisfied for $n > \tau_{2,\varepsilon}$. \square

LEMMA 6.8. *Suppose that Assumptions 2.1 – 2.3 hold. Then, there exists a random quantity \hat{C}_6 (which is a deterministic function of \hat{p} , \hat{C} , \hat{M}) such that the following is true: $1 \leq \hat{C}_6 < \infty$ everywhere and*

$$\liminf_{n \rightarrow \infty} \gamma_n^{\hat{p}} u(\theta_n) \leq \hat{C}_6 (\varphi_\varepsilon(\xi))^{\hat{\mu}} \quad (6.41)$$

on $\Lambda \setminus N_0$ for any $\varepsilon \in (0, \infty)$.

Proof. Let $\hat{C}_6 = \hat{C}_2 + \hat{C}_3^{\hat{p}}$. We prove (6.41) by contradiction: Assume that (6.41) is violated for some sample ω from $\Lambda \setminus N_0$ (notice that the formulas which follow in the proof correspond to this sample) and some real number $\varepsilon \in (0, \infty)$. Consequently, there exists $n_0 > \tau_{2,\varepsilon}$ (depending on ω, ε) such that

$$u(\theta_n) \geq \hat{C}_6 \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \quad (6.42)$$

for $n \geq n_0$. Let $\{n_k\}_{k \geq 0}$ be defined recursively by $n_k = a(n_{k-1}, \hat{t})$ for $k \geq 1$. In what follows in the proof, we consider separately the cases $\hat{\mu} < 2$ and $\hat{\mu} = 2$.

Case $\hat{\mu} < 2$: Due to (6.42), we have

$$v(\theta_{n_k}) \leq \hat{C}_6^{-1/\hat{p}} \gamma_{n_k}(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}}.$$

On the other side, Lemma 6.5 (relations (6.22), (6.25)) and (6.42) yield

$$v(\theta_{n_{k+1}}) - v(\theta_{n_k}) \geq (\hat{t}/\hat{C}_3)(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} \geq (1/\hat{C}_3)(\gamma_{n_{k+1}} - \gamma_{n_k})(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}}$$

for $k \geq 0$ (notice that $\hat{t} \geq \gamma_{n_{k+1}} - \gamma_{n_k}$). Therefore,

$$\begin{aligned} (1/\hat{C}_3)(\gamma_{n_k} - \gamma_{n_0})(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} &\leq \sum_{i=0}^{k-1} (v(\theta_{n_{i+1}}) - v(\theta_{n_i})) \\ &= v(\theta_{n_k}) - v(\theta_{n_0}) \\ &\leq \hat{C}_6^{-1/\hat{p}} \gamma_{n_k}(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} \end{aligned}$$

for $k \geq 1$. Thus,

$$(1 - \gamma_{n_0}/\gamma_{n_k}) \leq \hat{C}_3 \hat{C}_6^{-1/\hat{p}}$$

for $k \geq 1$. However, this is impossible, since the limit process $k \rightarrow \infty$ (applied to the previous relation) yields $\hat{C}_3 \geq \hat{C}_6^{1/\hat{p}}$ (notice that $\hat{C}_6 > \hat{C}_3^{\hat{p}}$). Hence, (6.41) holds when $\hat{\mu} < 2$.

Case $\hat{\mu} = 2$: As a result of Lemma 6.5 (relations (6.21), (6.24)) and (6.42), we get

$$u(\theta_{n_{k+1}}) \leq (1 - \hat{t}/\hat{C}_3)u(\theta_{n_k}) \leq \left(1 - (\gamma_{n_{k+1}} - \gamma_{n_k})/\hat{C}_3\right) u(\theta_{n_k})$$

for $k \geq 0$. Consequently,

$$\begin{aligned} u(\theta_{n_k}) &\leq u(\theta_{n_0}) \prod_{i=1}^k \left(1 - (\gamma_{n_i} - \gamma_{n_{i-1}})/\hat{C}_3\right) \\ &\leq u(\theta_{n_0}) \exp \left(- (1/\hat{C}_3) \sum_{i=1}^k (\gamma_{n_i} - \gamma_{n_{i-1}}) \right) \\ &= u(\theta_{n_0}) \exp \left(- (\gamma_{n_k} - \gamma_{n_0})/\hat{C}_3 \right) \end{aligned}$$

for $k \geq 0$. Then, (6.42) yields

$$\hat{C}_6(\varphi_\varepsilon(\xi))^{\hat{\mu}} \leq u(\theta_{n_0}) \gamma_{n_k}^{\hat{p}} \exp \left(- (\gamma_{n_k} - \gamma_{n_0})/\hat{C}_3 \right)$$

for $k \geq 0$. However, this is not possible, as the limit process $k \rightarrow \infty$ (applied to the previous relation) implies $\hat{C}_6(\varphi_\varepsilon(\xi))^{\hat{\mu}} \leq 0$. Thus, (6.41) holds also when $\hat{\mu} = 2$. \square

LEMMA 6.9. *Suppose that Assumptions 2.1 – 2.3 hold. Then, there exists a random quantity \hat{C}_7 (which is a deterministic function of \hat{p} , \hat{C} , \hat{M}) such that the following is true: $1 \leq \hat{C}_7 < \infty$ everywhere and*

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} u(\theta_n) \leq \hat{C}_7(\varphi_\varepsilon(\xi))^{\hat{\mu}} \quad (6.43)$$

on $\Lambda \setminus N_0$ for any $\varepsilon \in (0, \infty)$.

Proof. Let $\tilde{C}_1 = 3\hat{C}_1\hat{C}_5$, $\tilde{C}_2 = 6\tilde{C}_1\hat{C}_2 + \hat{C}_3^{\hat{p}} + \hat{C}_6$ and $\hat{C}_7 = 2(\tilde{C}_1 + \tilde{C}_2)^2$. We use contradiction to show (6.43): Suppose that (6.43) is violated for some sample ω from $\Lambda \setminus N_0$ (notice that the formulas which appear in the proof correspond to this sample) and some real number $\varepsilon \in (0, \infty)$. Then, it can be deduced from Lemma 6.8

that there exist $n_0 > m_0 > \tau_{2,\varepsilon}$ (depending on ω, ε) such that

$$\gamma_{m_0}^{\hat{p}} u(\theta_{m_0}) \leq \tilde{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}}, \quad (6.44)$$

$$\gamma_{n_0}^{\hat{p}} u(\theta_{n_0}) \geq \hat{C}_7(\varphi_\varepsilon(\xi))^{\hat{\mu}}, \quad (6.45)$$

$$\min_{m_0 < n \leq n_0} \gamma_n^{\hat{p}} u(\theta_n) > \tilde{C}_2(\varphi_\varepsilon(\xi))^{\hat{\mu}}, \quad (6.46)$$

$$\max_{m_0 \leq n < n_0} \gamma_n^{\hat{p}} u(\theta_n) < \hat{C}_7(\varphi_\varepsilon(\xi))^{\hat{\mu}} \quad (6.47)$$

(notice that $\tilde{C}_2 > \hat{C}_6$) and such that

$$(\gamma_{a(m_0, \hat{t})}/\gamma_{m_0})^{\hat{p}} \leq \min\{2, (1 - \hat{t}/\hat{C}_3)^{-1}\}, \quad (6.48)$$

$$\gamma_{m_0}^{-2r}(\xi + \varepsilon)^2 \leq \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \quad (6.49)$$

(to see that (6.48) holds for all, but finitely many m_0 , notice that $\lim_{n \rightarrow \infty} \gamma_{a(n, \hat{t})}/\gamma_n = 1$; to conclude that (6.49) is true for all, but finitely many m_0 , notice that $\hat{p} < 2 \min\{r, \hat{r}\} \leq 2r$ if $\hat{\mu} < 2$ and that the left and right-hand sides of (6.49) are equal when $\hat{\mu} = 2$).

Let $l_0 = a(m_0, \hat{t})$. As a direct consequence of Lemmas 6.2, 6.7 (relations (6.5), (6.40)) and (6.49), we get

$$\begin{aligned} u(\theta_n) - u(\theta_{m_0}) &\leq \hat{C}_1 (\gamma_{m_0}^{-r} \|\nabla f(\theta_{m_0})\|(\xi + \varepsilon) + \gamma_{m_0}^{-2r}(\xi + \varepsilon)^2) \\ &\leq \hat{C}_1 (\|\nabla f(\theta_{m_0})\|^2/2 + 3\gamma_{m_0}^{-2r}(\xi + \varepsilon)^2/2) \\ &\leq \hat{C}_1 \hat{C}_5 \psi(u(\theta_{m_0})) + (2\hat{C}_1 + \hat{C}_1 \hat{C}_5) \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \\ &\leq \tilde{C}_1 (\psi(u(\theta_{m_0})) + \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}}) \end{aligned} \quad (6.50)$$

for $m_0 \leq n \leq l_0$. Then, (6.46), (6.48), (6.50) yield

$$\begin{aligned} u(\theta_{m_0}) + \tilde{C}_1 \psi(u(\theta_{m_0})) &\geq u(\theta_{m_0+1}) - \tilde{C}_1 \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \\ &\geq (\tilde{C}_2 \gamma_{m_0+1}^{-\hat{p}} - \tilde{C}_1 \gamma_{m_0}^{-\hat{p}})(\varphi_\varepsilon(\xi))^{\hat{\mu}} \\ &= \left(\tilde{C}_2 (\gamma_{m_0+1}/\gamma_{m_0})^{-\hat{p}} - \tilde{C}_1 \right) \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \\ &\geq (\tilde{C}_2/2 - \tilde{C}_1) \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} > 0 \end{aligned} \quad (6.51)$$

(notice that $(\gamma_{m_0+1}/\gamma_{m_0})^{\hat{p}} \leq (\gamma_{l_0}/\gamma_{m_0})^{\hat{p}} \leq 2$; also notice that $\tilde{C}_2/2 \geq 3\tilde{C}_1$), while (6.44), (6.48), (6.50) imply

$$\begin{aligned} u(\theta_n) &\leq (1 + \tilde{C}_1)u(\theta_{m_0}) + \tilde{C}_1 \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \\ &\leq (\tilde{C}_1 + \tilde{C}_2 + \tilde{C}_1 \tilde{C}_2) \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \\ &< (\hat{C}_7/2) (\gamma_n/\gamma_{m_0})^{\hat{p}} \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \\ &\leq \hat{C}_7 \gamma_n^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \end{aligned} \quad (6.52)$$

for $m_0 \leq n \leq l_0$ (notice that $(\gamma_n/\gamma_{m_0})^{\hat{p}} \leq (\gamma_{l_0}/\gamma_{m_0})^{\hat{p}} \leq 2$ for $m_0 \leq n \leq l_0$; also notice that $\hat{C}_7/2 = (\tilde{C}_1 + \tilde{C}_2)^2 > \tilde{C}_1 + \tilde{C}_2 + \tilde{C}_1 \tilde{C}_2$). Due to (6.45), (6.47), (6.52), we have $l_0 < n_0$. On the other side, since $x + \tilde{C}_1 \psi(x) \geq 0$ only if $x \geq 0$ and since $x + \tilde{C}_1 \psi(x) = (1 + \tilde{C}_1)x$ for $x \geq 0$, inequality (6.51) implies

$$u(\theta_{m_0}) \geq (1 + \tilde{C}_1)^{-1} (\tilde{C}_2/2 - \tilde{C}_1) \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \geq \hat{C}_2 \gamma_{m_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \quad (6.53)$$

(notice that $\tilde{C}_2/2 - \tilde{C}_1 \geq \tilde{C}_1(3\hat{C}_2 - 1) \geq 2\tilde{C}_1\hat{C}_2 \geq (1 + \tilde{C}_1)\hat{C}_2$).

In what follows in the proof, we consider separately the cases $\hat{\mu} < 2$ and $\hat{\mu} = 2$.

Case $\hat{\mu} < 2$: Owing to Lemma 6.5 (relations (6.22), (6.25)) and (6.44), (6.53), we have

$$\begin{aligned} v(\theta_{l_0}) &\geq v(\theta_{m_0}) + (\hat{t}/\hat{C}_3)(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} \\ &\geq \left(\tilde{C}_2^{-1/\hat{p}}\gamma_{m_0} + \hat{C}_3^{-1}(\gamma_{l_0} - \gamma_{m_0}) \right) (\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} \\ &> \min\{\tilde{C}_2^{-1/\hat{p}}, \hat{C}_3^{-1}\}\gamma_{l_0}(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} \\ &= \tilde{C}_2^{-1/\hat{p}}\gamma_{l_0}(\varphi_\varepsilon(\xi))^{-\hat{\mu}/\hat{p}} \end{aligned}$$

(notice that $\hat{t} \geq \gamma_{l_0} - \gamma_{m_0}$; also notice $\tilde{C}_2^{-1/\hat{p}} < \hat{C}_3^{-1}$). Consequently,

$$u(\theta_{l_0}) = (v(\theta_{l_0}))^{-\hat{p}} < \tilde{C}_2\gamma_{l_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}}.$$

However, this directly contradicts (6.46) and the fact that $l_0 < n_0$. Thus, (6.43) holds when $\hat{\mu} < 2$.

Case $\hat{\mu} = 2$: Using Lemma 6.5 (relations (6.21), (6.24)) and (6.53), we get

$$u(\theta_{l_0}) \leq \left(1 - \hat{t}/\hat{C}_3\right) u(\theta_{m_0}).$$

Then, (6.44), (6.48) yield

$$u(\theta_{l_0}) \leq \tilde{C}_2(1 - \hat{t}/\hat{C}_3)(\gamma_{l_0}/\gamma_{m_0})^{\hat{p}}\gamma_{l_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \leq \tilde{C}_2\gamma_{l_0}^{-\hat{p}}(\varphi_\varepsilon(\xi))^{\hat{\mu}}.$$

However, this is impossible due to (6.46) and the fact that $l_0 < n_0$. Hence, (6.43) also in the case $\hat{\mu} = 2$. \square

LEMMA 6.10. *Suppose that Assumptions 2.1 – 2.3 hold. Then, for any real numbers $\varepsilon \in (0, \infty)$, $s \in (1, r] \cap (1, \hat{p})$, there exist a random quantity \hat{B}_s (which is a deterministic function of s , \hat{p} , \hat{C} , \hat{M} and does not depend on ε) and a non-negative integer-valued quantity $\sigma_{\varepsilon, s}$ such that the following is true: $1 \leq \hat{B}_s < \infty$, $0 \leq \sigma_{\varepsilon, s} < \infty$ everywhere and*

$$\sup_{k \geq n} \|\theta_k - \theta_n\| \leq \hat{B}_s\gamma_n^{-\hat{p}+s}(\varphi_\varepsilon(\xi))^{\hat{\mu}}(\phi_{\varepsilon, s}(\xi))^{-1} + \hat{B}_s\gamma_n^{-s+1}\phi_{\varepsilon, s}(\xi) \quad (6.54)$$

on $\Lambda \setminus N_0$ for $n > \sigma_{\varepsilon, s}$.

Proof. Let $\varepsilon \in (0, \infty)$, $s \in (1, r] \cap (1, \hat{p})$ be arbitrary real numbers, while $\tilde{C}_1 = 2(\hat{C}_2 + \hat{C}_7)$, $\tilde{C}_2 = 2\tilde{C}_1\hat{C}_5$, $\tilde{C}_3 = 2\tilde{C}_1\hat{C}_2$. Moreover, let

$$\tilde{B}_s = (3/\hat{t})(\tilde{C}_1 + \hat{C}_4)(1 + s/(\hat{p} - s) + 1/(s - 1))$$

and $\hat{B}_s = \tilde{B}_s + 2\tilde{C}_3$.

It is straightforward to show $\gamma_{a(n, \hat{t})} - \gamma_n = \hat{t} + O(\alpha_{a(n, \hat{t})})$ and

$$\begin{aligned} \gamma_{a(n, \hat{t})}^s - \gamma_n^s &= \gamma_{a(n, \hat{t})}^s \left(1 - \left(1 - (\gamma_{a(n, \hat{t})} - \gamma_n)/\gamma_{a(n, \hat{t})} \right)^s \right) \\ &= \gamma_{a(n, \hat{t})}^s \left(s\hat{t}\gamma_{a(n, \hat{t})}^{-1} + O(\gamma_{a(n, \hat{t})}^{-2}) \right) \end{aligned} \quad (6.55)$$

for $n \rightarrow \infty$. On the other side, Lemmas 6.7 and 6.9 imply

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} |u(\theta_n)| \leq \max\{\tilde{C}_2, \hat{C}_7\}(\varphi_\varepsilon(\xi))^{\hat{\mu}} \quad (6.56)$$

on $\Lambda \setminus N_0$, while Lemma 6.7 and (6.56) yield

$$\begin{aligned} \limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} \|\nabla f(\theta_n)\|^2 &\leq \hat{C}_5 \limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} \psi(u(\theta_n)) + \hat{C}_5 (\varphi_\varepsilon(\xi))^{\hat{\mu}} \\ &\leq 2\hat{C}_5 \max\{\hat{C}_2, \hat{C}_7\} (\varphi_\varepsilon(\xi))^{\hat{\mu}} \end{aligned} \quad (6.57)$$

on the same event. Then, owing to (6.55) – (6.57), it is possible to construct a non-negative integer-valued random quantity $\sigma_{\varepsilon,s}$ such that $\tau_{1,\varepsilon} \leq \sigma_{\varepsilon,s} < \infty$ everywhere and such that

$$\gamma_{a(n,\hat{t})} - \gamma_n \geq \hat{t}/2, \quad (6.58)$$

$$\gamma_{a(n,\hat{t})}^s - \gamma_n^s \leq s \gamma_{a(n,\hat{t})}^{s-1}, \quad (6.59)$$

$$|u(\theta_n)| \leq \tilde{C}_1 \gamma_n^{-\hat{p}} (\varphi_\varepsilon(\xi))^{\hat{\mu}}, \quad (6.60)$$

$$\|\nabla f(\theta_n)\| \leq \tilde{C}_2 \gamma_n^{-\hat{p}+s} (\varphi_\varepsilon(\xi))^{\hat{\mu}} (\phi_{\varepsilon,s}(\xi))^{-1} + \tilde{C}_2 \gamma_n^{-s+1} \phi_{\varepsilon,s}(\xi) \quad (6.61)$$

on $\Lambda \setminus N_0$ for $n > \sigma_{\varepsilon,s}$.⁵

Let ω is an arbitrary sample from $\Lambda \setminus N_0$ (notice that all formulas which follow in the proof correspond to this sample). Moreover, let $\{n_k\}_{k \geq 0}$ be recursively defined by $n_0 = \sigma_{\varepsilon,s} + 1$ and $n_{k+1} = a(n_k, \hat{t})$ for $k \geq 0$. Then, due to Lemma 6.6, we have

$$\begin{aligned} \|\theta_{n_l} - \theta_{n_k}\| &\leq \sum_{i=k}^{l-1} \|\theta_{n_{i+1}} - \theta_{n_i}\| \\ &\leq \sum_{i=k}^{l-1} \gamma_{n_i}^s (u(\theta_{n_i}) - u(\theta_{n_{i+1}})) (\phi_{\varepsilon,s}(\xi))^{-1} + \hat{C}_4 \sum_{i=k}^{l-1} \gamma_{n_i}^{-s} \phi_{\varepsilon,s}(\xi) \\ &\leq \sum_{i=k+1}^l (\gamma_{n_i}^s - \gamma_{n_{i-1}}^s) |u(\theta_{n_i})| (\phi_{\varepsilon,s}(\xi))^{-1} + \hat{C}_4 \sum_{i=k}^{l-1} \gamma_{n_i}^{-s} \phi_{\varepsilon,s}(\xi) \\ &\quad + \gamma_{n_l}^s |u(\theta_{n_l})| (\phi_{\varepsilon,s}(\xi))^{-1} + \gamma_{n_k}^s |u(\theta_{n_k})| (\phi_{\varepsilon,s}(\xi))^{-1} \end{aligned}$$

for $0 \leq k \leq l$. Consequently, (6.59), (6.60) yield

$$\begin{aligned} \|\theta_{n_l} - \theta_{n_k}\| &\leq \tilde{C}_1 s (\varphi_\varepsilon(\xi))^{\hat{\mu}} (\phi_{\varepsilon,s}(\xi))^{-1} \sum_{i=k+1}^l \gamma_{n_i}^{-\hat{p}+s-1} + \hat{C}_4 \phi_{\varepsilon,s}(\xi) \sum_{i=k}^{l-1} \gamma_{n_i}^{-s} \\ &\quad + \tilde{C}_1 (\gamma_{n_k}^{-\hat{p}+s} + \gamma_{n_l}^{-\hat{p}+s}) (\varphi_\varepsilon(\xi))^{\hat{\mu}} (\phi_{\varepsilon,s}(\xi))^{-1} \end{aligned} \quad (6.62)$$

for $0 \leq k \leq l$. Since

$$\gamma_{n_l} = \gamma_{n_k} + \sum_{i=k}^{l-1} (\gamma_{n_{i+1}} - \gamma_{n_i}) \geq \gamma_{n_k} + (\hat{t}/2)(l-k)$$

for $0 \leq k \leq l$ (owing to (6.58)), we get

$$\begin{aligned} \sum_{i=k}^{\infty} \gamma_{n_i}^{-\lambda-1} &\leq \sum_{i=0}^{\infty} (\gamma_{n_k} + i\hat{t}/2)^{-\lambda-1} \\ &\leq \gamma_{n_k}^{-\lambda-1} + \int_0^{\infty} (\gamma_{n_k} + u\hat{t}/2)^{-\lambda-1} du \\ &\leq 3\lambda^{-1} \hat{t}^{-1} \gamma_{n_k}^{-\lambda} \end{aligned}$$

⁵ To deduce that (6.61) holds on $\Lambda \setminus N_0$ for all but finitely many n , notice that $\|\nabla f(\theta_n)\| = O(\gamma_n^{-\hat{p}/2})$ on $\Lambda \setminus N_0$ and that at least one of $\hat{p}/2 > \hat{p} - s$, $\hat{p}/2 > s - 1$ is true.

for $k \geq 0$ and $\lambda \in (0, \infty)$. Then, (6.62) implies

$$\begin{aligned} \|\theta_{n_l} - \theta_{n_k}\| &\leq \tilde{C}_1 (2 + 3\hat{t}^{-1}s(\hat{p} - s)^{-1}) \gamma_{n_k}^{-\hat{p}+s}(\varphi_\varepsilon(\xi))^{\hat{\mu}}(\phi_{\varepsilon,s}(\xi))^{-1} \\ &\quad + 3\hat{C}_4 \hat{t}^{-1}(s-1)^{-1} \gamma_{n_k}^{-s+1} \phi_{\varepsilon,s}(\xi) \\ &\leq \tilde{B}_s \gamma_{n_k}^{-\hat{p}+s}(\varphi_\varepsilon(\xi))^{\hat{\mu}}(\phi_{\varepsilon,s}(\xi))^{-1} + \tilde{B}_s \gamma_{n_k}^{-s+1} \phi_{\varepsilon,s}(\xi) \end{aligned} \quad (6.63)$$

for $0 \leq k \leq l$. On the other side, since $s-1 < r$ and $\varphi_{\varepsilon,s}(\xi) \geq \xi + \varepsilon$, Lemma 6.2 (inequality (6.4)) and (6.61) yield

$$\begin{aligned} \|\theta_k - \theta_n\| &\leq \hat{C}_1 (\|\nabla f(\theta_n)\| + \gamma_n^{-r}(\xi + \varepsilon)) \\ &\leq \hat{C}_1 (\|\nabla f(\theta_n)\| + \gamma_n^{-s+1} \phi_{\varepsilon,s}(\xi)) \\ &\leq \tilde{C}_3 \gamma_n^{-\hat{p}+s}(\varphi_\varepsilon(\xi))^{\hat{\mu}}(\phi_{\varepsilon,s}(\xi))^{-1} + \tilde{C}_3 \gamma_n^{-s+1} \phi_{\varepsilon,s}(\xi) \end{aligned}$$

for $\sigma_{\varepsilon,s} < n \leq k \leq a(n, \hat{t})$ (notice that $\sigma_{\varepsilon,s} \geq \tau_{1,\varepsilon}$). Combining this with (6.63), we obtain

$$\begin{aligned} \|\theta_k - \theta_n\| &\leq \|\theta_k - \theta_{n_j}\| + \|\theta_{n_j} - \theta_{n_i}\| + \|\theta_{n_i} - \theta_n\| \\ &\leq \tilde{B}_s \gamma_{n_i}^{-\hat{p}+s}(\varphi_\varepsilon(\xi))^{\hat{\mu}}(\phi_{\varepsilon,s}(\xi))^{-1} + \tilde{B}_s \gamma_{n_i}^{-s+1} \phi_{\varepsilon,s}(\xi) \\ &\quad + \tilde{C}_3 (\gamma_n^{-\hat{p}+s} + \gamma_{n_j}^{-\hat{p}+s})(\varphi_\varepsilon(\xi))^{\hat{\mu}}(\phi_{\varepsilon,s}(\xi))^{-1} + \tilde{C}_3 (\gamma_n^{-s+1} + \gamma_{n_j}^{-s+1}) \phi_{\varepsilon,s}(\xi) \\ &\leq \hat{B}_s \gamma_n^{-\hat{p}+s}(\varphi_\varepsilon(\xi))^{\hat{\mu}}(\phi_{\varepsilon,s}(\xi))^{-1} + \hat{B}_s \gamma_n^{-s+1} \phi_{\varepsilon,s}(\xi) \end{aligned}$$

for $\sigma_{\varepsilon,s} < n \leq k$, $1 \leq i \leq j$ satisfying $n_{i-1} \leq n < n_i$, $n_j \leq k < n_{j+1}$. Then, it is obvious that (6.54) is true. \square

LEMMA 6.11. *Suppose that Assumptions 2.1 – 2.3 hold. Then, there exists a random quantity \hat{C}_8 (which is a deterministic function of $r, \hat{\mu}, \hat{C}, \hat{M}$) such that the following is true: $1 \leq \hat{C}_8 < \infty$ everywhere and such that*

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{q}} \sup_{k \geq n} \|\theta_k - \theta_n\| \leq \hat{C}_8 \varphi_\varepsilon(\xi) \quad (6.64)$$

on $\Lambda \setminus N_0$ for any $\varepsilon \in (0, \infty)$.

Proof. Let $\hat{s} = \min\{r, \hat{r}\}$ and $\hat{C}_8 = 2\hat{B}_{\hat{s}}$, while $\varepsilon \in (0, \infty)$ is an arbitrary real number. Moreover, let ω be an arbitrary sample from $\Lambda \setminus N_0$ (notice that all formulas which follow in the proof correspond to this sample). In order to show (6.64), we consider separately the cases $r < \hat{r}$ and $r \geq \hat{r}$.

Case $r < \hat{r}$: We have $\hat{s} = r$, $\hat{q} = r - 1$ and $\hat{p} = \hat{\mu}r = r(2 - 1/\hat{r})$. Consequently, $\hat{p} - r = r - r/\hat{r} > 0$, $\hat{p} - 2r + 1 = 1 - r/\hat{r} > 0$ (notice that $r/\hat{r} < 1 < r$), i.e., $\hat{s} < \hat{p}$, $\hat{q} < \hat{p} - \hat{s}$. Then, Lemma 6.10 implies

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{q}} \sup_{k \geq n} \|\theta_k - \theta_n\| \leq \hat{B}_{\hat{s}} \phi_{\varepsilon,\hat{s}}(\xi) \leq \hat{C}_8 \varphi_\varepsilon(\xi)$$

(notice that $\lim_{n \rightarrow \infty} \gamma_n^{-\hat{p}+\hat{q}+\hat{s}} = 0$ and $\phi_{\varepsilon,\hat{s}}(\xi) = \varphi_\varepsilon(\xi)$). Thus, (6.64) holds when $r < \hat{r}$.

Case $r \geq \hat{r}$: We have $\hat{s} = \hat{r}$, $\hat{q} = \hat{r} - 1$ and $\hat{p} = \hat{\mu}\hat{r} = 2\hat{r} - 1$. Therefore, $\hat{q} = \hat{p} - \hat{s}$ and $\hat{s} = (\hat{p} + 1)/2 < \hat{p}$ (notice that $\hat{p} > 1$). Then, Lemma 6.10 yields

$$\begin{aligned} \limsup_{n \rightarrow \infty} \gamma_n^{\hat{q}} \sup_{k \geq n} \|\theta_k - \theta_n\| &\leq \hat{B}_{\hat{s}} (\varphi_\varepsilon(\xi))^{\hat{\mu}}(\phi_{\varepsilon,\hat{s}}(\xi))^{-1} + \hat{B}_{\hat{s}} \phi_{\varepsilon,\hat{s}}(\xi) \\ &\leq \hat{B}_{\hat{s}} (\varphi_{\varepsilon,\hat{s}}(\xi))^{\hat{\mu}-1} + \hat{B}_{\hat{s}} \varphi_\varepsilon(\xi) \\ &\leq \hat{C}_8 \varphi_\varepsilon(\xi) \end{aligned}$$

(notice that $\phi_{\varepsilon, \hat{s}}(\xi) = \varphi_\varepsilon(\xi)$, since both $r = \hat{s}$, $r > \hat{r}$ cannot hold if $r \geq \hat{r}$; also notice that $\hat{\mu} - 1 \leq 1$ and that $\varphi_\varepsilon(\xi) \geq 1$ if $r \geq \hat{r}$). Hence, (6.64) is true when $r \geq \hat{r}$. \square

PROOF OF THEOREMS 2.1 AND 2.2. Owing to Lemmas 6.3 and 6.10, $\hat{\theta} = \lim_{n \rightarrow \infty} \theta_n$ exists and satisfies $\nabla f(\hat{\theta}) = 0$ on $\Lambda \setminus N_0$. Thus, Theorem 2.2 holds. In addition, we have $\hat{Q} \subseteq \{\theta \in \mathbb{R}^{d_\theta} : \|\theta - \hat{\theta}\| \leq \delta_{\hat{\theta}}\}$ on $\Lambda \setminus N_0$ ($\delta_{\hat{\theta}}$ is specified in Remark 2.1). Therefore, on $\Lambda \setminus N_0$, random quantities $\hat{\mu}$, \hat{p} , \hat{r} defined in this section coincide with $\hat{\mu}$, \hat{p} , \hat{r} specified in Theorem 2.2 (see Remark 2.1). Similarly, \hat{C} , \hat{M} introduced in this section are identical to $C_{\hat{\theta}}$, $M_{\hat{\theta}}$ (specified in Section 2) on $\Lambda \setminus N_0$. Thus, Theorem 2.1 is true.

Let $\hat{K} = 2\hat{C}_5(\hat{C}_5 + \hat{C}_7) + \hat{C}_8$. Then, Lemmas 6.5, 6.8 and the limit process $\varepsilon \rightarrow 0$ imply

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} |u(\theta_n)| \leq (\hat{C}_2 + \hat{C}_7)(\varphi(\xi))^{\hat{\mu}} \leq \hat{K}(\varphi(\xi))^{\hat{\mu}}$$

on $\Lambda \setminus N_0$. Consequently, Lemma 6.5 yields

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} \|\nabla f(\theta_n)\|^2 \leq \hat{C}_5(\varphi(\xi))^{\hat{\mu}} + \hat{C}_5 \limsup_{n \rightarrow \infty} \gamma_n^{\hat{p}} \psi(u(\theta_n)) \leq \hat{K}(\varphi(\xi))^{\hat{\mu}}$$

on $\Lambda \setminus N_0$. On the other side, using Lemma 6.11, we get

$$\limsup_{n \rightarrow \infty} \gamma_n^{\hat{q}} \|\theta_n - \hat{\theta}\| \leq \hat{C}_8 \varphi(\xi) \leq \hat{K} \varphi(\xi)$$

on $\Lambda \setminus N_0$. Hence, Theorem 2.2 holds, too. \square

REMARK 6.4. Owing to Lemma 6.10, $\{\theta_n\}_{n \geq 0}$ converges to $\hat{\theta}$ on $\Lambda \setminus N_0$ at the rate $O(\gamma_n^{-\min\{\hat{p}-s, s-1\}})$, where $s \in (1, r] \cap (1, \hat{p})$. It is straightforward to show

$$\hat{q} = \max_{s \in (1, r] \cap (1, \hat{p})} \min\{\hat{p} - s, s - 1\}.$$

This suggests that $O(\gamma_n^{-\hat{q}})$ is the tightest bound on the convergence rate of $\{\theta_n\}_{n \geq 0}$ which can be obtained by the arguments Lemmas 6.6, 6.10 and 6.11 are based on.

7. Proof of Theorem 3.1. The following notation is used in this section. For $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$, $E_{\theta, z}(\cdot)$ denotes $E(\cdot | \theta_0 = \theta, Z_0 = z)$. Moreover, let

$$\begin{aligned} \xi_n &= F(\theta_n, Z_{n+1}) - \nabla f(\theta_n), \\ \xi_{1,n} &= \tilde{F}(\theta_n, Z_{n+1}) - (\Pi \tilde{F})(\theta_n, Z_n), \\ \xi_{2,n} &= (\Pi \tilde{F})(\theta_n, Z_n) - (\Pi \tilde{F})(\theta_{n-1}, Z_n), \\ \xi_{3,n} &= -(\Pi \tilde{F})(\theta_n, Z_{n+1}) \end{aligned}$$

for $n \geq 1$. Then, it is obvious that algorithm (3.1) admits the form (2.1), while Assumption 3.2 yields

$$\begin{aligned} \sum_{i=n}^k \alpha_i \gamma_i^r \xi_i &= \sum_{i=n}^k \alpha_i \gamma_i^r \xi_{1,i} + \sum_{i=n}^k \alpha_i \gamma_i^r \xi_{2,i} - \sum_{i=n}^k (\alpha_i \gamma_i^r - \alpha_{i+1} \gamma_{i+1}^r) \xi_{3,i} \\ &\quad - \alpha_{k+1} \gamma_{k+1}^r \xi_{3,k} + \alpha_n \gamma_n^r \xi_{3,n-1} \end{aligned} \quad (7.1)$$

for $1 \leq n \leq k$.

LEMMA 7.1. Let Assumption 3.1 hold. Then, there exists a real number $s \in (0, 1)$ such that $\sum_{n=0}^{\infty} \alpha_n^{1+s} \gamma_n^r < \infty$.

Proof. Let $p = (2 + 2r)/(2 + r)$, $q = (2 + 2r)/r$, $s = (2 + r)/(2 + 2r)$. Then, using the Hölder inequality, we get

$$\sum_{n=0}^{\infty} \alpha_n^{1+s} \gamma_n^r = \sum_{n=1}^{\infty} (\alpha_n^2 \gamma_n^{2r})^{1/p} \left(\frac{\alpha_n}{\gamma_n^2} \right)^{1/q} \leq \left(\sum_{n=1}^{\infty} \alpha_n^2 \gamma_n^{2r} \right)^{1/p} \left(\sum_{n=1}^{\infty} \frac{\alpha_n}{\gamma_n^2} \right)^{1/q}.$$

Since $\gamma_{n+1}/\gamma_n = 1 + \alpha_n/\gamma_n = O(1)$ for $n \rightarrow \infty$ and

$$\sum_{n=1}^{\infty} \frac{\alpha_n}{\gamma_n^2} = \sum_{n=1}^{\infty} \frac{\gamma_{n+1} - \gamma_n}{\gamma_n^2} \leq \sum_{n=1}^{\infty} \left(\frac{\gamma_{n+1}}{\gamma_n} \right)^2 \int_{\gamma_n}^{\gamma_{n+1}} \frac{dt}{t^2} \leq \frac{1}{\gamma_1} \max_{n \geq 0} \left(\frac{\gamma_{n+1}}{\gamma_n} \right)^2,$$

it is obvious that $\sum_{n=0}^{\infty} \alpha_n^{1+s} \gamma_n^r$ converges. \square

PROOF OF THEOREM 3.1. Let $Q \subset \mathbb{R}^{d_\theta}$ be an arbitrary compact set, while $s \in (0, 1)$ is a real number such that $\sum_{n=0}^{\infty} \alpha_n^{1+s} \gamma_n^r < \infty$. Obviously, it is sufficient to show that $\sum_{n=0}^{\infty} \alpha_n \gamma_n^r \xi_n$ converges w.p.1 on $\bigcap_{n=0}^{\infty} \{\theta_n \in Q\}$.

Due to Assumption 3.1, we have

$$\begin{aligned} \alpha_{n-1}^s \alpha_n \gamma_n^r &= (1 + \alpha_{n-1}(\alpha_n^{-1} - \alpha_{n-1}^{-1}))^s \alpha_n^{1+s} \gamma_n^r = O(\alpha_n^{1+s} \gamma_n^r), \\ (\alpha_{n-1} - \alpha_n) \gamma_n^r &= (\alpha_n^{-1} - \alpha_{n-1}^{-1}) (1 + \alpha_{n-1}(\alpha_n^{-1} - \alpha_{n-1}^{-1})) \alpha_n^2 \gamma_n^r = O(\alpha_n^2 \gamma_n^r), \\ \alpha_n(\gamma_{n+1}^r - \gamma_n^r) &= \alpha_n \gamma_n^r ((1 + \alpha_n/\gamma_n)^r - 1) = \alpha_n \gamma_n^r (r\alpha_n/\gamma_n + o(\alpha_n/\gamma_n)) = o(\alpha_n^2 \gamma_n^r) \end{aligned}$$

as $n \rightarrow \infty$. Consequently,

$$\sum_{n=0}^{\infty} \alpha_n^s \alpha_{n+1} \gamma_{n+1}^r < \infty, \quad (7.2)$$

$$\sum_{n=0}^{\infty} |\alpha_n \gamma_n^r - \alpha_{n+1} \gamma_{n+1}^r| \leq \sum_{n=0}^{\infty} \alpha_n |\gamma_n^r - \gamma_{n+1}^r| + \sum_{n=0}^{\infty} |\alpha_n - \alpha_{n+1}| \gamma_{n+1}^r < \infty. \quad (7.3)$$

On the other side, as a result of Assumption 3.3, we get

$$\begin{aligned} E_{\theta,z} (\|\xi_{1,n}\|^2 I_{\{\tau_Q > n\}}) &\leq 2E_{\theta,z} (\varphi_{Q,s}^2(Z_{n+1}) I_{\{\tau_Q > n\}}) + 2E_{\theta,z} (\varphi_{Q,s}^2(Z_n) I_{\{\tau_Q > n-1\}}), \\ E_{\theta,z} (\|\xi_{2,n}\|^2 I_{\{\tau_Q > n\}}) &\leq E_{\theta,z} (\varphi_{Q,s}(Z_n) \|\theta_n - \theta_{n-1}\|^s I_{\{\tau_Q > n-1\}}) \\ &\leq \alpha_{n-1}^s E_{\theta,z} (\varphi_{Q,s}^2(Z_n) I_{\{\tau_Q > n-1\}}), \\ E_{\theta,z} (\|\xi_{3,n}\|^2 I_{\{\tau_Q > n\}}) &\leq E_{\theta,z} (\varphi_{Q,s}^2(Z_{n+1}) I_{\{\tau_Q > n\}}) \end{aligned}$$

for all $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$, $n \geq 1$. Then, Assumption 3.1 and (7.2) yield

$$\begin{aligned} E_{\theta,z} \left(\sum_{n=1}^{\infty} \alpha_n^2 \gamma_n^{2r} \|\xi_{1,n}\|^2 I_{\{\tau_Q > n\}} \right) &\leq 4 \left(\sum_{n=1}^{\infty} \alpha_n^2 \gamma_n^{2r} \right) \sup_{n \geq 0} E_{\theta,z} (\varphi_{Q,s}^2(Z_n) I_{\{\tau_Q \geq n\}}) < \infty, \\ E_{\theta,z} \left(\sum_{n=1}^{\infty} \alpha_n \gamma_n^r \|\xi_{2,n}\|^2 I_{\{\tau_Q > n\}} \right) &\leq \left(\sum_{n=1}^{\infty} \alpha_{n-1}^s \alpha_n \gamma_n^r \right) \sup_{n \geq 0} E_{\theta,z} (\varphi_{Q,s}^2(Z_n) I_{\{\tau_Q \geq n\}}) < \infty \end{aligned}$$

for any $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$, while (7.3) implies

$$\begin{aligned}
& E_{\theta,z} \left(\sum_{n=1}^{\infty} |\alpha_n \gamma_n^r - \alpha_{n+1} \gamma_{n+1}^r| \|\xi_{3,n}\| I_{\{\tau_Q > n\}} \right) \\
& \leq \left(\sum_{n=1}^{\infty} |\alpha_n \gamma_n^r - \alpha_{n+1} \gamma_{n+1}^r| \right) \sup_{n \geq 0} (E_{\theta,z} (\varphi_{Q,s}^2(Z_n) I_{\{\tau_Q \geq n\}}))^{1/2} < \infty, \\
& E_{\theta,z} \left(\sum_{n=1}^{\infty} \alpha_{n+1}^2 \gamma_{n+1}^{2r} \|\xi_{3,n}\|^2 I_{\{\tau_Q > n\}} \right) \\
& \leq \left(\sum_{n=1}^{\infty} \alpha_{n+1}^2 \gamma_{n+1}^{2r} \right) \sup_{n \geq 0} E_{\theta,z} (\varphi_{Q,s}^2(Z_n) I_{\{\tau_Q \geq n\}}) < \infty
\end{aligned}$$

for each $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$. Since

$$E_{\theta,z} (\xi_{1,n} I_{\{\tau_Q > n\}} | \mathcal{F}_n) = \left(E_{\theta,z} (\tilde{F}(\theta_n, Z_{n+1}) | \mathcal{F}_n) - (\Pi \tilde{F})(\theta_n, Z_n) \right) I_{\{\tau_Q > n\}} = 0$$

w.p.1 for every $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$, $n \geq 1$, it can be deduced easily that series

$$\sum_{n=1}^{\infty} \alpha_n \gamma_n^r \xi_{1,n}, \quad \sum_{n=1}^{\infty} \alpha_n \gamma_n^r \xi_{2,n}, \quad \sum_{n=1}^{\infty} (\alpha_n \gamma_n^r - \alpha_{n+1} \gamma_{n+1}^r) \xi_{3,n}$$

converge w.p.1 on $\bigcap_{n=0}^{\infty} \{\theta_n \in Q\}$, as well as that $\lim_{n \rightarrow \infty} \alpha_n \gamma_n^r \xi_{3,n-1} = 0$ w.p.1 on the same event. Owing to this and (7.1), we have that $\sum_{n=0}^{\infty} \alpha_n \gamma_n^r \xi_n$ converges w.p.1 on $\bigcap_{n=0}^{\infty} \{\theta_n \in Q\}$. \square

8. Proof of Theorems 4.1 and 4.2. In this section, we use the following notation. For $\theta \in \mathbb{R}^{d_\theta}$, $x \in \mathbb{R}^N$, $y \in \mathbb{R}$ and $z = [x^T \ y]^T$, let

$$F(\theta, z) = -(y - G_\theta(x))H_\theta(x),$$

while $Z_{n+1} = [X_n^T \ Y_n]^T$ for $n \geq 0$. With this notation, it is obvious that algorithm (4.1) admits the form (3.1).

PROOF OF THEOREM 4.1. Let $\theta = [a_1 \cdots a_M \ b_{1,1} \cdots b_{M,N}]^T \in \mathbb{R}^{d_\theta}$, while

$$\delta_\theta = \frac{\varepsilon}{2KLMN(1 + \|\theta\|)}$$

and $\hat{U}_\theta = \{\eta \in \mathbb{C}^{d_\theta} : \|\eta - \theta\| < \delta_\theta\}$ (ε is specified in Assumption 4.1). Moreover, for $\eta = [c_1 \cdots c_M \ d_{1,1} \cdots d_{M,N}]^T \in \mathbb{C}^{d_\theta}$, $x = [x_1 \cdots x_N]^T \in \mathbb{R}^N$, let

$$\begin{aligned}
\hat{G}_\eta(x) &= \sum_{i=1}^M c_i \hat{\psi} \left(\sum_{j=1}^N d_{i,j} x_j \right), \\
\hat{f}(\eta) &= \frac{1}{2} \int (y - \hat{G}_\eta(x))^2 \pi(dx, dy).
\end{aligned}$$

Then, we have

$$\left| \sum_{j=1}^N d_{i,j} x_j - \sum_{j=1}^N b_{i,j} x_j \right| \leq \sum_{j=1}^N |d_{i,j} - b_{i,j}| |x_j| \leq \delta_\theta L N < \varepsilon$$

for all $\eta = [c_1 \cdots c_M \ d_{1,1} \cdots d_{M,N}]^T \in \hat{U}_\theta$, $1 \leq i \leq M$ and each $x = [x_1 \cdots x_N]^T \in \mathbb{R}^N$ satisfying $\|x\| \leq L$. Consequently, Assumption 4.1 implies

$$\begin{aligned} & \left| \sum_{i=1}^M c_i \hat{\psi} \left(\sum_{j=1}^N d_{i,j} x_j \right) - \sum_{i=1}^M a_i \psi \left(\sum_{j=1}^N b_{i,j} x_j \right) \right| \\ & \leq \sum_{i=1}^M |c_i - a_i| \left| \hat{\psi} \left(\sum_{j=1}^N d_{i,j} x_j \right) \right| + \sum_{i=1}^M |a_i| \left| \hat{\psi} \left(\sum_{j=1}^N d_{i,j} x_j \right) - \hat{\psi} \left(\sum_{j=1}^N b_{i,j} x_j \right) \right| \\ & \leq \delta_\theta K M + K \sum_{i=1}^M |a_i| \left| \sum_{j=1}^N d_{i,j} x_j - \sum_{j=1}^N b_{i,j} x_j \right| \\ & \leq \delta_\theta K M + \delta_\theta K L M N \|\theta\| < \varepsilon \end{aligned}$$

for any $\eta = [c_1 \cdots c_M \ d_{1,1} \cdots d_{M,N}]^T \in \hat{U}_\theta$ and each $x = [x_1 \cdots x_N]^T \in \mathbb{R}^N$ satisfying $\|x\| \leq L$. Then, it can be deduced that for all $x \in \mathbb{R}^N$ satisfying $\|x\| \leq L$, $\hat{G}_\eta(x)$ is analytical in η on \hat{U}_θ . On the other side, Assumption 4.1 yields

$$\begin{aligned} |\hat{G}_\eta(x)| & \leq \sum_{i=1}^M |c_i| \left| \hat{\psi} \left(\sum_{j=1}^N d_{i,j} x_j \right) \right| \leq K M \|\eta\|, \\ \left| \frac{\partial}{\partial c_k} \hat{G}_\eta(x) \right| & = \left| \hat{\psi} \left(\sum_{j=1}^N d_{k,j} x_j \right) \right| \leq K, \\ \left| \frac{\partial}{\partial d_{k,l}} \hat{G}_\eta(x) \right| & = \left| \hat{\psi}' \left(\sum_{j=1}^N d_{k,j} x_j \right) c_{k,l} \right| \leq K L \|\eta\| \end{aligned}$$

for all $\eta = [c_1 \cdots c_M \ d_{1,1} \cdots d_{M,N}]^T \in \hat{U}_\theta$, $1 \leq k \leq M$, $1 \leq l \leq N$ and each $x = [x_1 \cdots x_N]^T \in \mathbb{R}^N$ satisfying $\|x\| \leq L$. Therefore,

$$\|\nabla_\eta \hat{G}_\eta(x)\| \leq K L M N (1 + \|\eta\|)$$

for any $\eta \in \hat{U}_\theta$ and each $x \in \mathbb{R}^N$ satisfying $\|x\| \leq L$. Thus,

$$\|\nabla_\eta (y - \hat{G}_\eta(x))^2\| = 2|y - \hat{G}_\eta(x)| \|\nabla_\eta \hat{G}_\eta(x)\| \leq 2K^2 L^2 M^2 N (1 + \|\eta\|)^2$$

for all $\eta \in \hat{U}_\theta$ and each $x \in \mathbb{R}^N$, $y \in \mathbb{R}$ satisfying $\|x\| \leq L$, $|y| \leq L$. Then, the dominated convergence theorem and Assumption 4.2 imply that $\hat{f}(\cdot)$ is differentiable on \hat{U}_θ . Consequently, $\hat{f}(\cdot)$ is analytical on \hat{U}_θ . Since $f(\theta) = \hat{f}(\theta)$ for all $\theta \in \mathbb{R}^{d_\theta}$, we conclude that $f(\cdot)$ is real-analytic on entire \mathbb{R}^{d_θ} . \square

PROOF OF THEOREM 4.2. As $\{Z_n\}_{n \geq 0}$ can be interpreted as a Markov chain whose transition kernel does not depend on $\{\theta_n\}_{n \geq 0}$, it is straightforward to show that Assumptions 3.2 and 3.3 hold. The theorem's assertion then follows directly from Theorem 3.1. \square

9. Proof of Theorems 5.1 and 5.2. In this section, we use the following notation. For $n \geq 0$, let

$$Z_n = [X_n^T \ Y_n \cdots Y_{n-M+1} \ \varepsilon_n \ \psi_n^T \cdots \varepsilon_{n-N+1} \ \psi_{n-N+1}^T]^T,$$

while $d_z = L + (M + N)(N + 1)$. For $\theta \in \Theta$, let $\varepsilon_0^\theta = \dots = \varepsilon_{-N+1}^\theta = 0$, $\psi_0^\theta = \dots = \psi_{-N+1}^\theta = 0$, while $\{\varepsilon_n^\theta\}_{n \geq 0}$, $\{\psi_n^\theta\}_{n \geq 0}$ are defined by the following recursion:

$$\begin{aligned}\phi_{n-1}^\theta &= [Y_{n-1} \dots Y_{n-M} \varepsilon_{n-1}^\theta \dots \varepsilon_{n-N}^\theta]^T, \\ \varepsilon_n^\theta &= Y_n - (\phi_{n-1}^\theta)^T \theta, \\ \psi_n^\theta &= \phi_{n-1}^\theta - [\psi_{n-1}^\theta \dots \psi_{n-N}^\theta] D \theta, \\ Z_n^\theta &= [X_n^T Y_n \dots Y_{n-M+1} \varepsilon_n^\theta (\psi_n^\theta)^T \dots \varepsilon_{n-N+1}^\theta (\psi_{n-N+1}^\theta)^T]^T, \quad n \geq 1.\end{aligned}$$

Then, it is straightforward to verify that $\{\varepsilon_n^\theta\}_{n \geq 0}$ satisfies the recursion (5.2), as well as that $\psi_n^\theta = \nabla_\theta \varepsilon_n^\theta$ for $n \geq 0$. Moreover, it can be deduced easily that there exist a matrix valued function $G_\theta : \Theta \rightarrow \mathbb{R}^{d_z \times d_z}$ and a matrix $H \in \mathbb{R}^{d_z \times L}$ with the following properties:

- (i) G_θ is linear in θ and its eigenvalues lie in $\{z \in \mathbb{C} : |z| < 1\}$ for each $\theta \in \Theta$.
- (ii) Equations

$$Z_{n+1}^\theta = G_\theta Z_n^\theta + H V_n, \quad Z_{n+1} = G_{\theta_n} Z_n + H V_n$$

hold for all $\theta \in \Theta$, $n \geq 0$.

The following notation is also used in this section. For $\theta \in \Theta$, $x \in \mathbb{R}^L$, $y_1, \dots, y_M \in \mathbb{R}$, $e_1, \dots, e_N \in \mathbb{R}$, $f_1, \dots, f_N \in \mathbb{R}^{d_\theta}$, and $z = [x^T y_1 \dots y_M e_1 f_1^T \dots e_N f_N^T]^T$, let

$$F(\theta, z) = f_1 e_1, \quad \phi(\xi) = e_1^2,$$

while

$$\Pi_\theta(z, B) = E(I_B(G_\theta z + H V_0))$$

for a Borel-measurable set B from \mathbb{R}^{d_z} . Then, it can be deduced easily that recursion (5.3) – (5.6) admits the form of the algorithm considered in Section 3. Furthermore, it can be shown that

$$(\Pi^n \phi)(\theta, 0) = E((\varepsilon_n^\theta)^2), \tag{9.1}$$

$$(\Pi^n F)(\theta, 0) = E(\psi_n^\theta \varepsilon_n^\theta) = \nabla_\theta (\Pi^n \phi)(\theta, 0) \tag{9.2}$$

for each $\theta \in \Theta$, $n \geq 0$.

PROOF OF THEOREM 5.1. Let $m = E(Y_0)$ and $r_k = r_{-k} = \text{Cov}(Y_0, Y_k)$ for $k \geq 0$, while

$$\varphi(\omega) = \sum_{k=-\infty}^{\infty} r_k e^{-i\omega k}$$

for $\omega \in [-\pi, \pi]$. Moreover, for $\theta \in \Theta$, $z \in \mathbb{C}$, let $C_\theta(z) = A_\theta(z)/B_\theta(z)$, while

$$\alpha_\theta = 1 + \max_{\omega \in [-\pi, \pi]} |A_\theta(e^{i\omega})|, \quad \beta_\theta = \min_{\omega \in [-\pi, \pi]} |B_\theta(e^{i\omega})|, \quad \delta_\theta = \frac{\beta_\theta}{4d_\theta \alpha_\theta}.$$

Obviously, $1 \leq \alpha_\theta < \infty$, $0 < \beta_\theta, \delta_\theta < \infty$ (notice that the zeros of $B_\theta(\cdot)$ are outside $\{z \in \mathbb{C} : |z| \leq 1\}$).

As $\sum_{k=0}^{\infty} r_k < \infty$, $|\varphi(\cdot)|$ is uniformly bounded. Consequently, the spectral theory for stationary processes (see e.g. [7, Chapter 2]) yields

$$\begin{aligned}\lim_{n \rightarrow \infty} E(\varepsilon_n^\theta) &= C_\theta(1)m, \\ \lim_{n \rightarrow \infty} \text{Cov}(\varepsilon_n^\theta, \varepsilon_{n+k}^\theta) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} |C_\theta(e^{i\omega})|^2 \varphi(\omega) e^{i\omega k} d\omega\end{aligned}$$

for all $\theta \in \Theta$, $k \geq 0$ (notice that $\varepsilon_n^\theta = C_\theta(q)Y_n$ and the poles of $C_\theta(\cdot)$ are in $\{z \in \mathbb{C} : |z| > 1\}$). Therefore,

$$f(\theta) = \frac{1}{4\pi} \int_{-\pi}^{\pi} |C_\theta(e^{i\omega})|^2 \varphi(\omega) d\omega + |C_\theta(1)|^2 \frac{m^2}{2} \quad (9.3)$$

for any $\theta \in \Theta$. On the other side, it is straightforward to verify

$$\begin{aligned}\frac{\partial}{\partial a_k} A_\theta(e^{i\omega}) &= -e^{-i\omega k}, \\ \frac{\partial^2}{\partial a_{k_1} \partial a_{k_2}} A_\theta(e^{i\omega}) &= 0, \\ \frac{\partial^{l_1+\dots+l_N}}{\partial b_1^{l_1} \dots \partial b_N^{l_N}} \left(\frac{1}{B_\theta(e^{i\omega})} \right) &= -(l_1 + l_2 + \dots + l_N)! e^{-i\omega(l_1+2l_2+\dots+Nl_N)} \\ &\quad \cdot \left(-\frac{1}{B_\theta(e^{i\omega})} \right)^{l_1+l_2+\dots+l_N+1}\end{aligned}$$

for every $\theta = [a_1 \dots a_M \ b_1 \dots b_N]^T \in \Theta$, $\omega \in [-\pi, \pi]$, $1 \leq k, k_1, k_2 \leq M$, $l_1, \dots, l_N \geq 0$. Thus,

$$\begin{aligned}& \left| \frac{\partial^{k_1+\dots+k_M+l_1+\dots+l_N}}{\partial a_1^{k_1} \dots \partial a_M^{k_M} \partial b_1^{l_1} \dots \partial b_N^{l_N}} C_\theta(e^{i\omega}) \right| \\ &= \left| \frac{\partial^{k_1+\dots+k_M}}{\partial a_1^{k_1} \dots \partial a_M^{k_M}} A_\theta(e^{i\omega}) \right| \left| \frac{\partial^{l_1+\dots+l_N}}{\partial b_1^{l_1} \dots \partial b_N^{l_N}} \left(\frac{1}{B_\theta(e^{i\omega})} \right) \right| \\ &\leq (l_1 + \dots + l_N)! \alpha_\theta (1/\beta_\theta)^{l_1+\dots+l_N+1}\end{aligned}$$

for all $\theta = [a_1 \dots a_M \ b_1 \dots b_N]^T \in \Theta$, $\omega \in [-\pi, \pi]$, $k_1, \dots, k_M \geq 0$, $l_1, \dots, l_N \geq 0$. Then, it can be deduced easily

$$\left| \frac{\partial^{k_1+\dots+k_{d_\theta}}}{\partial \vartheta_1^{k_1} \dots \partial \vartheta_{d_\theta}^{k_{d_\theta}}} C_\theta(e^{i\omega}) \right| \leq (k_1 + \dots + k_{d_\theta})! (\alpha_\theta / \beta_\theta)^{k_1+\dots+k_{d_\theta}+1}$$

for all $\theta \in \Theta$, $\omega \in [-\pi, \pi]$, $k_1, \dots, k_{d_\theta} \geq 0$ (ϑ_i denotes the i -th component of θ). Since

$$\begin{aligned}\frac{\partial^{k_1+\dots+k_{d_\theta}}}{\partial \vartheta_1^{k_1} \dots \partial \vartheta_{d_\theta}^{k_{d_\theta}}} |C_\theta(e^{i\omega})|^2 &= \sum_{j_1=0}^{k_1} \dots \sum_{j_{d_\theta}=0}^{k_{d_\theta}} \binom{k_1}{j_1} \dots \binom{k_{d_\theta}}{j_{d_\theta}} \frac{\partial^{j_1+\dots+j_{d_\theta}}}{\partial \vartheta_1^{j_1} \dots \partial \vartheta_{d_\theta}^{j_{d_\theta}}} C_\theta(e^{i\omega}) \\ &\quad \cdot \frac{\partial^{(k_1-j_1)+\dots+(k_{d_\theta}-j_{d_\theta})}}{\partial \vartheta_1^{k_1-j_1} \dots \partial \vartheta_{d_\theta}^{k_{d_\theta}-j_{d_\theta}}} C_\theta(e^{i\omega})\end{aligned}$$

for each $\theta \in \Theta$, $\omega \in [-\pi, \pi]$, $k_1, \dots, k_{d_\theta} \geq 0$, we have

$$\begin{aligned}
& \left| \frac{\partial^{k_1+\dots+k_{d_\theta}}}{\partial \vartheta_1^{k_1} \dots \partial \vartheta_{d_\theta}^{k_{d_\theta}}} |C_\theta(e^{i\omega})|^2 \right| \\
& \leq (k_1 + \dots + k_{d_\theta})! \left(\frac{\alpha_\theta}{\beta_\theta} \right)^{k_1+\dots+k_{d_\theta}+2} \sum_{j_1=0}^{k_1} \dots \sum_{j_{d_\theta}=0}^{k_{d_\theta}} \frac{\binom{k_1}{j_1} \dots \binom{k_{d_\theta}}{j_{d_\theta}}}{\binom{k_1+\dots+k_{d_\theta}}{j_1+\dots+j_{d_\theta}}} \\
& \leq (k_1 + \dots + k_{d_\theta})! \left(\frac{\alpha_\theta}{\beta_\theta} \right)^{k_1+\dots+k_{d_\theta}+2} \sum_{j_1=0}^{k_1} \dots \sum_{j_{d_\theta}=0}^{k_{d_\theta}} \binom{k_1}{j_1} \dots \binom{k_{d_\theta}}{j_{d_\theta}} \\
& \leq (k_1 + \dots + k_{d_\theta})! \left(\frac{2\alpha_\theta}{\beta_\theta} \right)^{k_1+\dots+k_{d_\theta}+2}
\end{aligned}$$

for any $\theta \in \Theta$, $\omega \in [-\pi, \pi]$, $k_1, \dots, k_{d_\theta} \geq 0$. Consequently, the multinomial formula (see [12, Theorem 1.3.1]) implies

$$\begin{aligned}
& \sum_{k_1=0}^{\infty} \dots \sum_{k_{d_\theta}=0}^{\infty} \frac{\delta_\theta^{k_1+\dots+k_{d_\theta}}}{k_1! \dots k_{d_\theta}!} \left| \frac{\partial^{k_1+\dots+k_{d_\theta}}}{\partial \vartheta_1^{k_1} \dots \partial \vartheta_{d_\theta}^{k_{d_\theta}}} |C_\theta(e^{i\omega})|^2 \right| \\
& \leq \left(\frac{2\alpha_\theta}{\beta_\theta} \right)^2 \sum_{k_1=0}^{\infty} \dots \sum_{k_{d_\theta}=0}^{\infty} \frac{(k_1 + \dots + k_{d_\theta})!}{k_1! \dots k_{d_\theta}!} \left(\frac{2\alpha_\theta \delta_\theta}{\beta_\theta} \right)^{k_1+\dots+k_{d_\theta}} \\
& = \left(\frac{2\alpha_\theta}{\beta_\theta} \right)^2 \sum_{n=0}^{\infty} \sum_{\substack{0 \leq k_1, \dots, k_{d_\theta} \leq n \\ k_1+\dots+k_{d_\theta}=n}} \frac{(k_1 + \dots + k_{d_\theta})!}{k_1! \dots k_{d_\theta}!} \left(\frac{2\alpha_\theta \delta_\theta}{\beta_\theta} \right)^{k_1+\dots+k_{d_\theta}} \\
& = \left(\frac{2\alpha_\theta}{\beta_\theta} \right)^2 \sum_{n=0}^{\infty} \left(\frac{2d_\theta \alpha_\theta \delta_\theta}{\beta_\theta} \right)^n \\
& = \left(\frac{2\alpha_\theta}{\beta_\theta} \right)^2 \sum_{n=0}^{\infty} \left(\frac{1}{2} \right)^n < \infty
\end{aligned}$$

for every $\theta \in \Theta$, $\omega \in [-\pi, \pi]$. Then, the analyticity of $f(\cdot)$ directly follows from (9.3) and the fact that $|\varphi(\cdot)|$ is uniformly bounded (also notice that $C_\theta(1)$ is analytic in θ). \square

PROOF OF THEOREM 5.2. It is straightforward to show

$$\begin{aligned}
& \max\{\|F(\theta, z)\|, \|\phi(z)\|\} \leq \|z\|, \\
& \max\{\|F(\theta, z') - F(\theta, z'')\|, |\phi(z') - \phi(z'')|\} \leq 2\|z' - z''\|(\|z'\| + \|z''\|)
\end{aligned}$$

for all $\theta \in \Theta$, $z, z', z'' \in \mathbb{R}^{d_z}$. Moreover, it can be deduced easily that for any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exist real numbers $\delta_{1,Q} \in (0, 1)$, $C_{1,Q} \in [1, \infty)$ such that $\|G_\theta^n\| \leq C_{1,Q} \delta_{1,Q}^n$ and

$$\|G_{\theta'} - G_{\theta''}\| \leq C_{1,Q} \|\theta' - \theta''\|$$

for each $\theta, \theta', \theta'' \in Q$, $n \geq 0$. Then, the results of [1, Section II.2.3] imply that there exist a locally Lipschitz continuous function $g : \Theta \rightarrow \mathbb{R}^{d_\theta}$ and a Borel-measurable function $\tilde{F} : \Theta \times \mathbb{R}^{d_z} \rightarrow \mathbb{R}^{d_\theta}$ such that

$$F(\theta, z) - g(\theta) = \tilde{F}(\theta, z) - (\Pi \tilde{F})(\theta, z)$$

for every $\theta \in \Theta$, $z \in \mathbb{R}^{d_z}$. Due to the same results, there exists a locally Lipschitz continuous function $h : \Theta \rightarrow \mathbb{R}$ and for any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exist real numbers $\delta_{2,Q} \in (0, 1)$, $C_{2,Q} \in [1, \infty)$ such that

$$\begin{aligned} \max\{\|(\Pi^n F)(\theta, z) - g(\theta)\|, \|(\Pi^n \phi)(\theta, z) - h(\theta)\|\} &\leq C_{2,Q} \delta_{2,Q}^n (1 + \|z\|)^2, \\ \max\{\|\tilde{F}(\theta, z)\|, \|(\Pi \tilde{F})(\theta, z)\|\} &\leq C_{2,Q} (1 + \|z\|)^2, \\ \|\tilde{F}(\theta', z) - \tilde{F}(\theta'', z)\| &\leq C_{2,Q} \|\theta' - \theta''\| (1 + \|z\|)^2 \end{aligned} \quad (9.4)$$

for each $\theta, \theta', \theta'' \in Q$, $z, z', z'' \in \mathbb{R}^{d_z}$. Combining (9.1), (9.2), (9.4) with the dominated convergence theorem, we get $h(\cdot) = f(\cdot)$, $g(\cdot) = \nabla f(\cdot)$. On the other side, owing to the fact that $\{X_n\}_{n \geq 0}$ is a geometrically ergodic Markov chain, we have that $\{Y_n\}_{n \geq 0}$ admits a stationary regime for $n \rightarrow \infty$. Consequently, Theorem 5.1 implies that $f(\cdot)$ is analytic on Θ . Then, the theorem's assertion directly follows from Theorem 3.1. \square

Appendix. In this section, we prove the claim stated in Remark 2.2. If open set V specified in Remark 2.2 exists, we can define the following quantities for any compact set $Q \subset \mathbb{R}^{d_\theta}$ and any $a \in f(Q)$:

$$\begin{aligned} \tilde{\delta}_{Q,a} &= \begin{cases} \delta_{\tilde{Q},a}, & \text{if } Q \cap S \neq \emptyset, a \in f(S) \\ 1, & \text{if } Q \cap S = \emptyset \\ \min\{1, d(a, f(S))/2\}, & \text{if } a \notin f(S) \end{cases} \\ \tilde{\mu}_{Q,a} &= \begin{cases} \mu_{\tilde{Q},a}, & \text{if } Q \cap S \neq \emptyset, a \in f(S) \\ 2, & \text{otherwise} \end{cases} \\ \tilde{M}_{Q,a} &= 1 + \sup \left\{ \frac{|f(\theta) - a|}{\|\nabla f(\theta)\|^{\tilde{\mu}_{Q,a}}} : \theta \in Q \setminus S, |f(\theta) - a| \leq \tilde{\delta}_{Q,a} \right\} \end{aligned}$$

where $\tilde{Q} = Q$ if $Q \subset V$ and $\tilde{Q} = \{\theta \in Q : d(\theta, S) \leq d(Q \setminus V, S)/2\}$ otherwise. Then, it is straightforward to show

$$\begin{aligned} a \notin f(S) &\implies \inf\{\|\nabla f(\theta)\| : \theta \in Q, |f(\theta) - a| \leq \tilde{\delta}_{Q,a}\} > 0, \\ Q \setminus V \neq \emptyset &\implies \inf\{\|\nabla f(\theta)\| : \theta \in Q \setminus \tilde{Q}\} > 0, \\ Q \cap S \neq \emptyset &\implies \sup \left\{ \frac{|f(\theta) - a|}{\|\nabla f(\theta)\|^{\tilde{\mu}_{Q,a}}} : \theta \in \tilde{Q}, |f(\theta) - a| \leq \tilde{\delta}_{Q,a} \right\} \leq M_{\tilde{Q},a} < \infty. \end{aligned}$$

Consequently, $\tilde{\delta}_{Q,a}$, $\tilde{\mu}_{Q,a}$, $\tilde{M}_{Q,a}$ are well-defined and enjoy the following properties: $0 < \tilde{\delta}_{Q,a} \leq 1$, $1 < \tilde{\mu}_{Q,a} \leq 2$, $1 \leq \tilde{M}_{Q,a} < \infty$ and

$$|f(\theta) - a| \leq \tilde{M}_{Q,a} \|\nabla f(\theta)\|^{\tilde{\mu}_{Q,a}}$$

for all $\theta \in Q$ satisfying $|f(\theta) - a| \leq \tilde{\delta}_{Q,a}$. Hence, the claim holds.

REFERENCES

- [1] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, 1990.
- [2] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.
- [3] D. P. Bertsekas, *Nonlinear Programming*, 2nd edition, Athena Scientific, 1999.

- [4] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM Journal on Optimization, 10 (2000), pp. 627 – 642.
- [5] E. Bierstone and P. D. Milman, *Semianalytic and subanalytic sets*, Institut des Hautes Études Scientifiques, Publications Mathématiques, 67 (1988), pp. 5 - 42.
- [6] V. S. Borkar and S. P. Meyn, *The ODE Method for Convergence of Stochastic Approximation and Reinforcement Learning*, SIAM Journal on Control and Optimization, 38 (2000), pp. 447 – 469.
- [7] P. E. Caines, *Linear Stochastic Systems*, Wiley, 1988.
- [8] H.-F. Chen, *Stochastic Approximation and Its Application*, Kluwer, 2002.
- [9] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, Wiley, 2002.
- [10] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, 2001.
- [11] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, 1998.
- [12] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*, Birkhäuser, 2002.
- [13] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Annales de l’Institut Fourier (Grenoble), 48 (1998), pp. 769 - 783.
- [14] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd edition, Springer-Verlag, 2003.
- [15] L. Ljung, *Analysis of a general recursive prediction error identification algorithm*, Automatica, 27 (1981), pp. 89 – 100.
- [16] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*, MIT Press, 1983.
- [17] L. Ljung, *System Identification: Theory for the User*, 2nd edition, Prentice Hall, 1999.
- [18] S. Łojasiewicz, *Sur le problème de la division*, Studia Mathematica, 18 (1959), pp. 87 – 136.
- [19] S. Łojasiewicz, *Sur la géométrie semi- et sous-analytique*, Annales de l’Institut Fourier (Grenoble), 43 (1993), pp. 1575 – 1595.
- [20] M. Metivier and P. Priouret, *Applications of a Kushner-Clark lemma to general classes of stochastic algorithms*, IEEE Transactions on Information Theory, 30 (1984), pp. 140 – 151.
- [21] A. Nedić and D. P. Bertsekas, *Convergence Rate of Incremental Subgradient Algorithms*, in S. Uryasev and P. M. Pardalos (Eds.), *Stochastic Optimization: Algorithms and Applications*, Kluwer, pp. 263 – 304.
- [22] G. Ch. Pflug, *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*, Kluwer 1996.
- [23] B. T. Polyak and Y. Z. Tsypkin, *Criterion algorithms of stochastic optimization*, Automation and Remote Control, 45 (1984), pp. 766 – 774.
- [24] B. T. Polyak, *Introduction to Optimization*, Optimization Software, 1987.
- [25] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, Wiley, 2007.
- [26] J. C. Spall, *Introduction to Stochastic Search and Optimization*, Wiley, 2003.
- [27] V. B. Tadić, *On the Almost Sure Rate of Convergence of Linear Stochastic Approximation*, IEEE Transactions on Information Theory, 50 (2004), pp. 401 – 409.
- [28] V. B. Tadić, *Convergence Rate of Stochastic Gradient Search in the Case of Multiple and Non-Isolated Minima*, extended version of this paper, available at [arXiv:0904.4229v2](https://arxiv.org/abs/0904.4229v2).
- [29] V. B. Tadić, *Analyticity, Convergence and Convergence Rate of Recursive Maximum Likelihood Estimation in Hidden Markov Models*, submitted, available at [arXiv:0904.4264v1](https://arxiv.org/abs/0904.4264v1).